Secondary implementation of interactive engagement teaching techniques: Choices and challenges in a Gulf Arab context

G. W. Hitt, A. F. Isakovic, and O. Fawwaz
Department of Applied Mathematics and Sciences, Khalifa University of Science, Technology and Research, Abu Dhabi Campus, P. O. Box 127788, Abu Dhabi, United Arab Emirates

M. S. Bawa’anleh, N. El-Kork, S. Makkiyl, and I. A. Qattan
Department of Applied Mathematics and Sciences, Khalifa University of Science, Technology and Research, Sharjah Campus, P. O. Box 573, Sharjah, United Arab Emirates

We report on efforts to design the “Collaborative Workshop Physics” (CWP) instructional strategy to deliver the first interactive engagement (IE) calculus-based physics course at Khalifa University of Science, Technology and Research (KU), United Arab Emirates (UAE). To these authors’ knowledge, this work reports the first calculus-based physics course on the Arabian Peninsula using Physics Education Research (PER)-based instruction. A brief history and present context of general university and science/engineering teaching in the UAE is given. From this frame, a pre-reform baseline is presented for KU’s calculus-based introductory mechanics course in its traditional, lecture(er)-centered form, as established by Force Concept Inventory (FCI) and Maryland Physics Expectations in Physics (MPEX) survey data, as well as course exam grades. The first semester of delivery in the prototype CWP modality has identified several key challenges for further reforms: (1) second-language acquisition, (2) secondary and post-secondary academic preparation, and (3) culture-specific gender issues. We identify and characterize each of these issues quantitatively through analysis of pre/post survey, course exam and standardized test data. We find that for students with high English language proficiency, normalized gain on FCI improves substantially, from \( g = 0.16 \pm 0.10 \) pre-reform to \( g = 0.47 \pm 0.08 \) in the CWP pilot (standard errors). However, we also find evidence that normalized gains on FCI are strongly modulated by language proficiency and discuss likely causes. Regardless of language ability, problem-solving skill is also substantially improved and course drop-fail-withdrawal (DFW) rates are cut from 50% to 24%. Open questions are identified and recommendations are made for future improvements, relevant to KU, universities in the broader Gulf region and other institutions in the developing world facing similar challenges involving secondary implementation of PER-based instructional strategies.

I. INTRODUCTION

The use of interactive-engagement (IE) instructional strategies and curriculum resources developed through physics education research (PER) in North America and Europe has produced improved student problem-solving performance and deeper conceptual understanding relative to lecture-centered instruction (e.g. in introductory mechanics). More recently, increased attention has been given to the complications, and their mitigation, arising during secondary implementations of PER-based curricula (e.g. and to institutionalizing successful PER-based reforms. Specifically, evidence presented in Refs. shows that the broader contexts in which an interactive-engagement course is implemented is, for the success and sustainability of the implementation, at least as important as how well PER-based learning tasks are executed in the classroom. These broader contexts can include the departmental, institutional, student and faculty idio-/ethno-cultural contexts. Several broad research questions are raised, given these demonstrations of the importance of context. Specifically, how far away from the context of the developing institution can a PER-based instructional strategy be implemented? If one of the broader contexts mentioned above is very different to that of the original developing institution, are there criteria on these contexts that can help faculty who are planning a secondary implementation to predict possible risks for their reform project? Following from this, in terms of the implementation, how and to what extent can the original instructional strategy be changed in anticipation of these failure risks, so as to better match the contexts of the implementing institution, without compromising that instructional strategy’s core functions? Perhaps most importantly, is there a generic change strategy that faculty groups can follow to help them achieve success and sustainability for their efforts across all of the relevant contexts, especially departmental and institutional contexts?

The present work makes contributions toward answering these questions, by reporting on a modified implementation of cooperative group problem-solving (CGPS) in a United Arab Emirates (UAE) context at Khalifa University of Science, Technology, and Research (KU hereafter). Motivated by Refs., this work also presents a design-based approach for choosing and changing the CGPS instructional strategy based on an analysis of the cultural expectations of its users (students), and presents a post-analysis of efficacy. Our long term vision at KU is to address major questions related to secondary
implementations for the UAE and the broader Arabian Peninsula/Gulf Arab context. The first step toward this goal is to answer certain narrower, more concrete questions for the KU/UAE context which are:

1. **On the equivalence of lecture-centered approaches:** Is there a correspondence between features and effects of traditional, lecture-centered instruction in the US and those of lecture-centered instruction in the UAE? Do lecture-centered approaches to introductory, calculus-based physics in the two societies share the same features, in terms of classroom expectations and norms, instructional approaches and curriculum content, and the effects of the instruction on conceptual learning, problem-solving skill and course drop-fail-withdrawal (DFW) rates?

2. **On the equivalence of IE approaches:** Does use of the CGPS approach, thoughtfully modified, produce improvements in student conceptual learning and problem-solving ability in KU students that are similar to comparable secondary implementations presented in the PER literature?

3. **On identification and mitigation of failure risks for secondary implementations:** Are there features of KU/UAE contexts that threaten successful and sustainable implementation of CGPS and how can these risks be mitigated? Are these features similar to those faced by secondary implementations conducted within the US or are there qualitatively different challenges? What changes to the CGPS approach are suggested by these differences and to what extent can an instructional strategy be adapted to them while maintaining its basic integrity?

To answer these questions, this work is structured as follows. In Sec. II a brief overview of the UAE and KU contexts will be given, starting with a historical perspective and a summary of the present state of affairs, with emphasis on the role and perception of higher education in UAE society. The section focuses on a variety of measures of student values and expectations, of learning in UAE society. The section will give emphasis on the role and perception of higher education contexts will be given, starting with a historical perspective.

**II. UAE AND KU CONTEXTS**

In this section, we briefly review the major and relatively recent historical developments in the Gulf region, as they relate to education, for the benefit of the reader and to inform the methodology and discussion of this study.

**A. A Brief History of Education in the UAE**

Major political and economic changes in the MENA region often initiate or come in tandem with large-scale educational reforms (see Refs. 16,17 for further review). Education in the lower Gulf coast of the Arabian Peninsula is no exception and has undergone several rapid changes in recent history, first with the pearling industry boom of the late 19th century, with that industry's collapse during the Great Depression and the World Wars, and with the discovery of oil in the middle of the 20th century. Prior to the arrival of European colonial empires, the regional economy was mostly subsistence and did not permit the labor specialization necessary for formal education. Rather, in small, informal gatherings a mutawwa'a, a respected community elder who was often the worship leader of the local mosque, would lead neighborhood boys in Islamic religious oral recitations and teach general wisdom for life.

During the middle decades of the 19th century, the British Empire entered into trade and security agreements with the coastal sheikhdoms in an effort to secure sea lanes to India. These "Trucial States" brought peace and, combined with warm, shallow seas, a commercial pearling boom that in turn funded the first formal schools. Tragically, the truce agreements that brought prosperity also forbade Gulf Arab merchants from trading pearls outside imperial markets. In the 1920's and 30's, when Europe was hit by depression and war and Japan introduced cultivated pearls, the Gulf pearling industry collapsed completely and formal schooling all but disappeared. The ensuing hardships and lack of access to education persisted even after the discovery of oil in the Trucial States territory in October, 1969.

On December 2, 1971, following British withdrawal, the seven hereditary monarchies of Abu Dhabi, Ajman, Dubai, Fujairah, Ras al-Khaimah, Sharjah, and Umm al-Qaiwain declared their formation of the United Arab Emirates. Concurrently, the UAE Ministry of Education, along with many other federal ministries, were created to
oversee a new public school system, using curricula imported mainly from Kuwait and Jordan and primarily teacher-driven, rote-learning methods, as textbooks and other resources were not widely available. This quickly changed as the oil crises of the 1970’s, combined with rapid growth in worldwide oil use, lead to huge expansions in affluence and access to education for the region. Over the span of a few generations, the society rapidly transformed from one of about 80,000 Gulf Arabs, with a per capita income of 3K USD (2005 dollars) and an adult literacy rate of < 10%, to one that at present has nearly 6 million people, with expatriate groups from 90 nations, a per capita income of 33K USD, and an adult literacy rate amongst citizens of ~80%.

At present, there are 19 institutions of tertiary education in the Emirate of Abu Dhabi alone, including KU. A few salient features of the current higher education landscape are as follows. Combined, these institutions have a gross enrollment of about 25-30% of the adult citizen population, a factor of 5 increase over that of 1970 and 75% of which are female students. The language of instruction in most settings is English. Consequently, while each institution has a distinct core mission, all that teach in English share a need to accommodate a majority of English Language Learners (ELLs), graduating from the mostly Arabic-based secondary schools. For these students, most institutions have a language-conditional admission category and a “foundation” or “preparatory” program (see a similar example in Ref.23), a year-long, intensive English and remedial math-and-science curriculum. As a result, the average time spent studying to obtain a bachelor’s degree is 5.5 years. Another ongoing challenge, especially for STEM-focused programs, is the relatively small number of students following science and mathematics-intensive tracks in secondary school (< 5%) and selecting to study STEM disciplines at university (< 30%). KU was established in 2008 by royal decree and by acquisition of Etisalat University College in Sharjah, UAE which now forms its Sharjah campus facility. The Abu Dhabi campus opened its doors to degree program students in Fall 2009. Currently, the University is composed of the College of Engineering only which includes the Department of Applied Mathematics and Sciences where its mathematics and natural sciences faculty are employed. All students, about 1000 enrolled total, are engineering majors who take two calculus-based introductory physics courses delivered by the department. Across the two campuses, these two courses currently serve 250-300 students per year, about 75% of whom are UAE citizens.

B. Student expectations from national culture and transfer of norms across contexts

Figure 1 shows a schematic representation of some US and UAE teaching and learning contexts that are relevant for the secondary implementation of PER-based instructional strategies, created largely in the US, to their parallel frames of context in UAE society. Vertical arrows represent ways in which norms in one frame inform norms in their contained frames. Horizontal arrows represent ways in which norms in one society might be transferred to parallel frames in the other society. As described by Finkelstein4,5. We add horizontal arrows to represent ways in which norms in one frame inform norms in their contained frames, as described by Finkelstein4,5. For ease of reference, we typify the transfers by frame. Type I is
transfer of norms at the task level (e.g. introducing the expectation that students be able to solve novel or purely conceptual problems). Type II is transfer of situational norms (e.g. introducing the expectation that students interact with each other during class time). Type III is the transfer of idio-cultural norms (e.g. introducing the expectation that university “life” means high student autonomy and self-regulation). Type IV is the transfer of national/ethno-cultural norms (e.g. introducing the expectation that one queue for a teller at the bank in a line).

Transfers of the Type IV kind are highly unlikely in the short term (and attempting them is often highly unproductive), but faculty awareness of the differences in expectations about university, between US national and UAE national culture, is important for student motivational issues. A US professor’s motivational speech usually goes something like ‘study hard because you’re paying for tuition and if you do well, you’ll make more money when you graduate’ which is mostly meaningless in the UAE context. As a society where domestic tertiary education is relatively new, mostly imported, and entirely subsidized, UAE citizens have significantly different views about the role of higher education in their society as compared to US citizens and consequently UAE students attend university for different reasons. While Arab Gulf states primarily establish universities to enlarge and diversify their private sector economies away from oil and reduce unemployment, individuals and families in these nations view universities as a means to gain social status, increase their marriageability, and secure public sector employment. Essentially all Arab Gulf states offer their citizens access to tuition-free tertiary education and, unless they pursue other forms of employment, most graduates desire and have guaranteed employment with their governments. Amongst degree-holding Omanis, Qataris, and Emirati citizens, 86%, 87%, and 85% are employed in the public sector, respectively. This is in sharp contrast to US students, who largely attend university to broaden their career choices and increase their lifetime earning potential. Aside from awareness, it’s beyond the scope of this paper to discuss Type IV transfers further, so we focus on Types I - III for the rest of the text.

We further narrow our scope by assuming that, at the task level, the average student is cognitively the same regardless of culture and that Type I transfers (of published PER-based tasks and task-level norms) must and should be successful without major modification. The key issue then is effective Type II and Type III transfers, those of the situational and idio-cultural norms critically linked to effective use of PER-based tasks. To maximize the likelihood of a successful and sustainable secondary implementation, this should be done with minimization of negative expectancy violation (for students and instructors alike) as a key goal in the course design. There is evidence that Gulf Arab university students’ expectations for classroom norms may not be as incompatible with the idio-cultural values (e.g. student-student interactions, hands-on activities, equality in teacher-student interactions, sense-making over answer-making) intrinsic to successful IE pedagogies, contrary to what casual observers of UAE national/ethno-culture might imagine.

In fact, there is an interesting juxtaposition present in Fig. 1. The seminal work on cultural values by G. Hofstede would suggest that American culture, when manifest in the educational situational/idio-cultural frames, carries expectations of/preferences for: student-teacher equality, student-centered education, students initiating communication, open-ended learning situations, good discussions, tasks with uncertain outcomes that involve risk estimation and problem solving, grouping according to tasks, speaking out in class or in large groups, encouragement for showing individual initiative, and emphasis learning ‘how to learn’. These expectations all appear very much in-line with those of functional IE pedagogy, but Hofstede cautions that factors such as affluence, education, occupation, gender and age can significantly affect these expectations. Indeed, as one looks into narrower frames of context for the US, a large body of research on expectations for introductory physics courses shows these are in fact, not consistent with the expectations of American university students. One could speculate at the cause. Perhaps because values implicit in most US public primary and secondary education are at odds with broader societal and family values in the US and as a result, IE-compatible habits-of-mind (e.g. divergent thinking) are devalued and decline with prolonged formal education. Traditional instruction at the university level would then be expected by American students as a result of prior experience rather than on culture, and continuing that instructional mode continues the trend away from expert-like attitudes toward the course content.

Conversely, the same considerations applied to Gulf Arab students suggests that as one looks into narrower frames of context on the right side of Fig. 1 these trends are reversed as compared to the US scenario. Hofstede’s analysis of broader Gulf Arab culture suggests that in educational situational/idio-cultural frames, students expect/prefer: student deference and dependence on teachers, teacher-centered education, teacher-initiated communication, highly structured learning situations, teachers possessing and transferring absolute truths, tasks with sure outcomes that involve following instructions and no risk, grouping according to prior affiliations (ethnicity, family, friendship, etc.), no speaking out in class or in large groups, discouragement for individual initiative, and emphasis on learning ‘how to learn’. These expectations all appear at odds with IE classroom norms, but when Hofstede’s survey instrument (VSM94; see Ref.) is given specifically to Gulf Arab university students in their classroom environments by their teachers (i.e. precisely in the relevant frame of context), there are significant modifications. J. Baron surveyed 200 pre-
health sciences students at the University of Sharjah (USHj) and many student classroom expectations (those associated with power distance and uncertainty avoidance) were similarly or better aligned with IE pedagogy than one would expect of general American culture. In a survey of 219 foundation/preparatory program students at Qatar University (QU), K. Litvin and M. McAllister similarly found that students expected greater equality with their teachers and were more comfortable with uncertainty in learning tasks and environments than an analysis of the broader culture would suggest.

C. Student idio-cultural expectations at KU

Anecdotally, these authors find that many of the classroom behaviors exhibited by KU students are consistent with those reported by Litvin and McAllister, especially their observations on student reading habits, preparations prior to class time/lecture, and communication habits. In both institutions, students coming directly from high school have always had all materials (paper, pens, books, etc.) provided for them on a same-day basis and rarely realize that they need to bring or take notes during class, so these behaviors have to be taught and reinforced. Also, Gulf Arab culture has a strong oral tradition and reading for pleasure is quite rare, especially since their has never been a wide variety of books available until very recently. As a result, students at QU and KU rarely read their course textbooks unless there are direct and regular grade incentives to do so. Furthermore, while students generally do prefer interactive and group learning, they do still prefer to work with family, tribe, or close friends in class and on occasion, teaming a given student with certain other students can lead to tense situations, as was also observed at QU.

Perhaps most importantly, similar to QU, these authors find that while KU students do expect student-centeredness and student equality in most regards, there are important ways in which the deference to authority, that is expected in the broader culture, manifests in the university setting, particularly around 'non-curricular' issues like grievances about grades, classroom management, assignment deadlines, etc. KU students will rarely confront a faculty member personally with a request for special treatment (e.g. permission to make up an assignment due to illness, absence for travel, etc.) or with a complaint (e.g. that pace of lecture is too fast, grading of a previous assignment unfair, etc.). Instead, students will communicate in ways so that the issue is presented as one belonging to a third party, which appears to make such conversations more passive and more comfortable for them. For special requests, students will often send a friend to speak with the faculty on their behalf. The student’s friend finds it easier to make a special request of the faculty since the request is 'not for me' and the student themself likely feels that they are being more submissive and respectful of the teacher’s authority by not asking them 'to their face'. Even more importantly, for complaints, students may instead speak directly to the faculty’s supervisor or other high-level administrator, again arranging the discussion about the issue so that it is about someone who is not personally present and thereby avoids direct confrontation and offense to the teacher.

This behavior can have very serious effects on a course if not anticipated by all parties involved, especially since US trained faculty and administrators generally consider student complaints that 'go over a faculty’s head’ (such as to an ombudsman) as indicative of serious faculty abuse. If faculty and administrators are unfamiliar or misinformed about Gulf Arab culture, they may not entirely grasp what the reasonable expectations of students are in a given situation and misjudge the seriousness of a complaint. Students also soon realize that 'cultural offense', and the staff uncertainty and anxiety it produces, can be a convenient “Trojan Horse” for getting any complaint to be taken seriously. Indeed, during the 2009-2011 period, there were multiple examples of essentially instant changes to the management of courses that were prompted by relatively minor complaints, often to the surprise of faculty and students alike. This experience made it clear to these authors that any pedagogical reform project would have to anticipate the nature of student expectancy violation and prepare project evaluators and administrators in advance, to expect complaints and to take care separating justifiable from unjustifiable ones. Likewise, to increase the likelihood of success, faculty would have to 'pick their battles' and avoid expectancy violation that is pedagogically unnecessary (e.g. issues of Type IV transfer of norms discussed above).

There are also some significant differences between student expectations at USHj (as studied in Ref.) and QU (as studied in Ref.) and those at KU. The most important one is that the students are separated by gender at QU and USHj, and they are not separated at KU. Classroom geometry in KU classrooms typically has left-right symmetry and a large center isle, communicating unconsciously to students that the genders should sit on opposite sides which they invariably do. So, while the classroom spaces are co-educational, this is not without some tension, and in a 2011 town hall discussion with students, many students expressed a desire that learning activities be further separated along gender lines. Specifically, some students (from both genders) felt more comfortable addressing faculty when members of the opposite sex were not present and that they could better learn subject-matter content in single-gender settings, with gender integration used only for team-based projects. Consequently, teaming male and female students in groups during class must be done very thoughtfully, if done at all. It is very likely that doing so may not produce learning gains that are worth the discomfort produced, or worse, that learning gains could be undermined by lack of student engagement caused by their discomfort.
intrinsic expectations. Note also that compared to the MPEX
in the CWP pilot (2011) are not caused by differences
in expectations. Figure 2 shows the overall score on the MPEX, adminis-
ter pre-instruction, and displayed in an agree-disagree
(A-D) plot as done by Reddish et al. (1998). Overall,
both year groups test similarly (errors in the mean are
only a few percent, smaller than the icons used in the
plot). Thus, there is some confidence that differences be-
tween traditionally taught students (2010) and students
in the CWP pilot (2011) are not caused by differences in
expectations. Note also that compared to the MPEX
validation group and student groups surveyed in Red-
dish et al.’s original publication (“US Tertiary” in the
figure), KU students respond significantly less favorably
on the MPEX pre-test. In fact, KU pre-test scores on
MPEX are consistent with US
on the MPEX pre-test. In fact, KU pre-test scores on

D. Students’ task-level expectations for learning
physics content

In this section, we are concerned with students expec-
tations of physics learning itself, at the narrowest frame,
the task-level context. To gauge students in this regard,
we have surveyed two groups, one from academic year
2010 and one from the pilot CWP group in 2011, using
the Maryland Physics Expectation (MPEX) survey12.

Figure 2 shows the A-D plot for the independence clus-
ter. Clearly, KU students respond very unfavorably to
some items in this cluster. The most unfavorable re-

sponses are to items #1 and #14, as was the case in the
original MPEX study12, however the degree of unfavora-
bility is significantly greater. Item #1 states:

All I need to do to understand most of the
basic ideas in this course is just read the text,
work most of the problems, and/or pay close
attention in class.

Only 8% of students on average (9% for year group 2010,
7% for 2011) disagreed with this statement (the favorable
response). No student strongly disagreed. On aver-
arge 77% of students agreed with item #1 (76% for KU
2010 and 78% for KU 2011), meaning on average only
15% of students responded neutrally. A similar pattern
is present in responses to item #14 which states:

Learning physics is a matter of acquiring
knowledge that is specifically located in the
laws, principles, and equations given in class
and/or in the textbook.

Only 17% of students on average (18% for KU 2010, 14% for
KU 2011) disagreed with this statement (the favorable
response). Again, no student strongly disagreed. On aver-
age 56% of students agreed with item #14 (59% for KU
2010, 50% for KU 2011), meaning on average 27% of stu-
dents responded neutrally. On the remaining items in the
independence cluster, students on average responded fa-
vorably 39% versus 35% unfavorably and 26% neutrally,
still slightly lower but somewhat more consistent with
US tertiary pre-test scores presented in Ref.12. (Note: Cronbach’s alpha score for the independence cluster is
0.58).

This stands somewhat in contrast to cultural values
on similar students discussed in the previous section
and helps to triangulate on some of the anecdotal remarks
about KU students that were added there. Students do
expect student-centeredness and student-teacher equality,
but this appears to be mainly when they are engaging
the instructor of the course. When engaging the content
of the course, clearly, based on the text of item #1 and
#14, KU students exhibit a particularly high dependence
on perceived authoritative sources of information (text,
problems, or ‘in class’ [instructor]) as compared to US
students on average.

As shown in Fig. 1, the effort cluster from the MPEX
data presents a more favorable, though not entirely clear
picture. Clearly, KU students respond favorably to most
items in this cluster, much more like their US tertiary
counterparts prior to instruction. The notable and some-
what confusing exceptions are items #6 and #24 which
received on average 48% and 35% favorable responses,
respectively. Item #6 states:

Only 8% of students on average (9% for year group 2010,
7% for 2011) disagreed with this statement (the favorable
response). No student strongly disagreed. On aver-
average 77% of students agreed with item #1 (76% for KU
2010 and 78% for KU 2011), meaning on average only
15% of students responded neutrally. A similar pattern
is present in responses to item #14 which states:

Learning physics is a matter of acquiring
knowledge that is specifically located in the
laws, principles, and equations given in class
and/or in the textbook.

Only 17% of students on average (18% for KU 2010, 14% for
KU 2011) disagreed with this statement (the favorable
response). Again, no student strongly disagreed. On aver-
age 56% of students agreed with item #14 (59% for KU
2010, 50% for KU 2011), meaning on average 27% of stu-
dents responded neutrally. On the remaining items in the
independence cluster, students on average responded fa-
vorably 39% versus 35% unfavorably and 26% neutrally,
still slightly lower but somewhat more consistent with
US tertiary pre-test scores presented in Ref.12. (Note: Cronbach’s alpha score for the independence cluster is
0.58).

This stands somewhat in contrast to cultural values
on similar students discussed in the previous section
and helps to triangulate on some of the anecdotal remarks
about KU students that were added there. Students do
expect student-centeredness and student-teacher equality,
but this appears to be mainly when they are engaging
the instructor of the course. When engaging the content
of the course, clearly, based on the text of item #1 and
#14, KU students exhibit a particularly high dependence
on perceived authoritative sources of information (text,
problems, or ‘in class’ [instructor]) as compared to US
students on average.

As shown in Fig. 1, the effort cluster from the MPEX
data presents a more favorable, though not entirely clear
picture. Clearly, KU students respond favorably to most
items in this cluster, much more like their US tertiary
counterparts prior to instruction. The notable and some-
what confusing exceptions are items #6 and #24 which
received on average 48% and 35% favorable responses,
respectively. Item #6 states:
I spend a lot of time figuring out and understanding at least some of the derivations or proofs given either in class or in the text.

A 48% agreement (the favorable response) indicates that either proofs themselves, or the contexts in which they are given (in-class or the text) is not deemed an important behavior for learning physics. Yet, item #7 states:

I read the text in detail and work through many of the examples given there.

The level of agreement (the favorable response) is 70% on average. Taken together, it would appear that KU students believe that engagement with the course textbook is important for learning physics, but it is most valuable as a source of worked examples, not as a source of proofs and derivations. This is somewhat in contradiction to the instructors’ anecdotal evidence, discussed above, that many students do not read the course text at all. It might be the case however, that the discrepancy is the result of the wording of item #6 which contains many compound constructions (“figuring out and understanding”, “derivations or proofs”, “in class or in the text”), that ELL students may not be sure what the statement is asking them to reflect upon. (Note: In support of this possibility, the Cronbach’s alpha score for the effort cluster is 0.60, but if item #6 is removed, it improves the most for any 1-item removal, to 0.70).

Item #24 is also responded to relatively unfavorably, with only 35% agreeing. It states:

The results of an exam don’t give me any useful guidance to improve my understanding of the course material. All the learning associated with an exam is in the studying I do before it takes place.

This response is strangely at odds with student responses to item #31:

I use the mistakes I make on homework and on exam problems as clues to what I need to do to understand the material better.

to which 83% of students agree, the most favorable of any in the effort cluster. Speculating, this discrepancy could be the result of the long tradition in UAE public schools to follow a British-style model for testing, where learning is assessed by a single, high-stakes exam at the end of the course (often carrying 60% or more of the course’s grade weight), as opposed to the model in typical US physics courses which feature 2-3 midterm exams and weekly graded homework spread throughout the course. In the prior case, with the course completed, there would be little reason for a student to expect to need to study their mistakes on an exam since the course is finished. Confirming this however, requires a more specific investigation. Nevertheless, the fact that 83% of students agree to the statement of item #31 and expect to study their mistakes on homework assignments in order to understand the course material is encouraging.

E. Summary Discussion of Context and Expectations

Student expectations associated with classroom norms at the various levels of context have been assessed. Data on UAE nationals’ cultural values and how these inform expectations in educational environments suggest that essential classroom norms for effective PER-based instruction will not be totally rejected. On the contrary, at the ethno/national cultural level, there is some evidence that course reform might be better received by students at KU than among students at US institutions. At the idio-cultural level, despite the benefits reported in the literature on cooperative, mixed-gender, female-majority teams (e.g. Ref 9), it seems unlikely in the KU context that the student learning gains amongst fresh-
III. BENCHMARKING TRADITIONAL INSTRUCTION

KU’s original model for science and engineering courses with labs is a so-called “3+1” inclusive model, with 3 credit hours per week devoted to lectures (and therefore 3 contact hours of lecture) and 1 credit hour devoted to a laboratory session which typically meets once per week for 2.5 to 3 contact hours to create a 4-credit hour course. The lecture and laboratory portion of the course form a single whole and students cannot register for one without the other. Laboratory grades contribute a percentage toward the overall weight of coursework and are added to contributions from exams to calculate a single course grade. Lectures in the first-semester calculus-based physics course cover material typically included in introductory mechanics: linear, circular, and rotational kinematics, forces and torques, Newton’s Laws, work, energy, conservation laws for energy, linear momentum, and angular momentum, statics and some basic topics in thermal physics. The original format of the course meets all of Hake’s criteria for classification as a “traditional” course; didactic lectures to a passive student audience, “recipe” or “verification” labs, and all formative and summative assessment with individual student exams featuring exclusively algorithmic problems (alternatively called end-of-chapter (e.g see references therein) or “Halliday-Resnick problems). Readers will notice the noteworthy absence of recitation or tutorial/problem-solving sessions. A typical semester would see a total of 45 lectures, delivered either in three 50-minute session per week or two 75-minute sessions, for 15 weeks and approximately 10 laboratory sessions spaced over the calendar to roughly follow the lecture and avoid weeks with holidays. As is often the case in many such courses in other universities, keeping the lecture and lab portions of the course synchronized is difficult and the topics being treated in the laboratory could often be ahead or behind the lecture by as much as two weeks.

There are also several features of the University’s learning environment that are unique, owing to its status as a recent start-up institution, that are worth mention. First, all students are granted full scholarship as part of their admission to the University. This is done to attract top talent and a large talent pool to aid in the start-up phase. Consequently, there are certain student traits, related to socio-economic background and personal motivation, that are difficult to compare with student bodies in many other universities. Second, as there is yet no dormitory facilities on the Abu Dhabi campus, most students are commuter students in one form or another. Access to campus is restricted in evening hours and there are no 24-hour facilities that are available to students. This necessarily means that some portion of their out-of-class studying occurs off campus which will be relevant when language issues are discussed below (see Sec. V).

FIG. 5. Histogram of pre- (black) and post-test (white) Force Concept Inventory data. a.) FCI data from all traditionally taught students. b.) Data from all traditionally taught, directly admitted students. c.) Data from all traditionally taught, conditionally admitted students. d.) Data from all conditionally admitted students taught in the CWP pilot.
A. Pre-Reform Baseline Performance Assessment

Starting from Fall 2009, Force Concept Inventory (FCI)\cite{13} was administered to students both pre-/post-instruction, in English. This was repeated in two subsequent semesters, Spring 2010 and Fall 2010, where the traditional course was offered however, logistical issues prevented administration of the Spring 2010 FCI post-test. In Fall of 2010 and Spring 2011, student had a choice of taking FCI in Arabic or English but this had no significant effect on average scores, pre- or post-instruction, over Fall 2009 or Spring 2010. Figure 5 shows the distribution of all FCI data considered in this work. Table I shows class-averaged pre- and post-instruction FCI scores for all semesters that data have been gathered. A strict matching condition has been applied, such that any student that did not complete both pre- and post-tests is removed from the dataset prior to analysis. The uncertainties quoted for pre-test, and post-test scores are errors in the mean ($\sqrt{\sigma^2/N}$). The uncertainties for the Hake’s gain scores are propagated from the uncertainties for mean pre-test and mean post-test scores, where the Hake’s gain score itself is calculated in the usual way as 

$$\langle g \rangle_{\text{Hake}} = \frac{\langle \text{Post-test} \rangle - \langle \text{Pre-test} \rangle}{100 - \langle \text{Pre-test} \rangle}$$

and where $\langle \rangle$ symbols indicate class-averaged scores.

B. Combining all data from traditional instruction on the AD Campus

Given the start-up status of the university, class sizes for the Fall 2009 offering of the course were quite small and it was some time before enough data could be gathered to begin making sound decisions about potential reform approaches. Near the end of Fall 2010, all data gathered at that time were culled together and analyzed to see if the data sets over the three traditional offerings (Fall 2009, Spring 2010, and Fall 2010) on the AD campus could be merged and analyzed as a single data set. With small sample sizes, it is difficult to confidently verify the normality of the data set, so we used a non-parametric test, the Wilcoxon (or Mann-Whitney) Rank Sum test\cite{14} (hereafter simply ‘rank-sum’ test), to compare means and distributions. Comparing FCI pre-test scores for these three offerings, we find the following Z-statistics for the rank-sum test as applied pair-wise to the three data sets: Fall 2009 and Spring 2010, $Z = -0.90$ and two-tailed p-value $p \sim 0.367$; Fall 2009 and Fall 2010, $Z = -0.02$ and two-tailed p-value $p \sim 0.985$; and Fall 2010 and Spring 2010, $Z = 1.18$ and two-tailed p-value $p \sim 0.239$. Since all of the two-tailed p-values are much greater than 0.05, we therefore conclude there are no significant differences in the FCI pre-test scores for students in these three offerings. This comparison was repeated using a parametric test, the Welch’s t-test, just as a check on this conclusion, even though it is difficult to verify normality in any of these individual data sets. The results for the two-tailed p-values are similarly large and likewise suggest that there are no statistically significant differences between the three sets of FCI pre-tests.

C. Pre-instruction evidence of a heterogeneous student population

Pre-test data gathered from Fall 2009 until Spring 2011 on the AD campus indicate that the student body is unusually heterogeneous and can be consistently decomposed into two major groups based on their admissions type: (1) direct-entry students (those directly admitted to credit-bearing, degree-granting programs) and (2) preparatory-track students. The main criteria for determining along which path a student enters are their performance on an English language equivalency test (IELTS\cite{15}, most often or occasionally TOEFL IBT\cite{16}), their secondary school percentile, performance on a pre-admissions battery of tests (testing basic mathematics knowledge and computer literacy), and performance in an on-site interview. This two-component feature of our student body is important to characterize and monitor quantitatively since, as a start-up institution, teaching across all subjects and at all levels is in a constant state of experimentation and flux and without monitoring it can be very difficult to attribute cause for improved learning to particular pedagogical changes. In this section, we initially focus on the data set from the Abu Dhabi campus, since the sample sizes are much larger, to answer some basic research questions. Later, we will justify merging the data sets, after checking if they are distributed similarly, and analyzing them jointly.

Coincidentally, all CWP pilot students on both campuses were, as indicated in Tab. I admitted to the University through the preparatory program however, it is important to determine how changes to preparatory program teaching (which includes conceptual and algebraic physics courses) during 2010 may have altered their pre-test performance compared to previous generations of preparatory-track students. Table II shows FCI pre-test scores, descriptive statistics, and rank-sum Z-statistics for comparison of direct-entry and prep populations and the CWP pilot students on the Abu Dhabi campus. Students repeating the course (and have therefore been re-administered the FCI multiple times previously) are removed from the data set, however we do not enforce strict matching here (i.e. its not necessary that a student have an FCI post-test to be included here) since we are not calculating or comparing gains. The rank-sum test is used to compare average scores on the various pre-assessments with the intent of determining if dissimilarities between the populations are statistically significant. The rank-sum test is an appropriate statistical test to use initially as it is a non-parametric test and can be used to analyze data sets that are discreet and have an unknown or indetermi-
nate distribution (e.g. negative exponential, Gaussian, etc.). Since we have removed students repeating the course, we satisfy the necessary precondition for using the rank-sum test, namely that the compared data sets are independent.

As shown in lower half of Tab. 11, the average FCI pre-test score of direct-entry students is significantly different from that of preparatory students or of students in the CWP pilot. As an example, comparing the direct-entry students with preparatory students, we take the null hypothesis to be that their respective mean FCI pre-test scores are equal. As shown in Tab. 11 the test statistic $Z$ for this comparison is 4.59 which has a corresponding two-tailed $p$-value of less than 0.00001. We can therefore confidently reject the null hypothesis. FCI pre-test scores for direct-entry students are 11% higher on average than those of preparatory students and the difference is statistically significant. We also find that FCI pre-test scores of direct-entry students are likewise higher and significantly different from those of our CWP pilot students ($p < 0.001$). When comparing preparatory students to the CWP pilot students however, we find that the null hypothesis, that their mean FCI pre-test scores are equal, cannot be confidently rejected ($p \sim 0.66$). We therefore can conclude that, despite changes to preparatory physics courses and pedagogy after Fall 2009 but prior to the CWP pilot, the FCI pre-test scores show no statistically significant differences.

As stated above, English language equivalency is a major factor deciding whether a student is admitted directly to a degree program or admitted conditionally to the preparatory program. Also, one of the major changes to the preparatory program prior to the launch of the CWP pilot was that the required IELTS score for direct admission was raised from an IELTS overall score of 5.5 to an overall score of 6.0. Therefore, as part of our characterization and monitoring of students’ pre-instruction traits, we analyze IELTS overall scores to determine if there are significant sub-populations based on language ability.

Table 11 shows IELTS overall scores, descriptive statistics, and rank-sum $Z$-statistics for comparison of direct-entry and prep populations and the CWP pilot students. Like the above analysis of FCI pre-test scores, students repeating the course are removed from the data set to avoid double-counting their IELTS overall scores (unlike FCI, they do not retake IELTS when repeating the physics course) and invalidating the use of the rank-sum test which is again used to compare the different populations by admissions type. The main reason the rank-sum is the appropriate statistical test to use here is that IELTS scores are discreet and very coarse (coming in 0.5 increments over a range of 4.0 to 9.0) and applying normality and parametric tests to small IELTS data sets is rather uninformative, as they often trivially pass such tests. As shown in the lower half of Tab. 11, the average IELTS overall score of direct-entry students is significantly different from that of preparatory students or of students in the CWP pilot (two-tailed $p < 0.001$). The difference between average IELTS scores for CWP pilot students and other preparatory students is not as pronounced but is still significant ($p < 0.001$). Other preparatory students have more broadly distributed IELTS scores while CWP IELTS scores peak sharply at 6.0 (more than 50% of CWP students have IELTS 6.0).

There is one last quantitative measure of the difference between the direct and preparatory admitted students.
TABLE II. Some average pre-test scores for direct entry, prep-track and CWP pilot student populations on the Abu Dhabi campus.

| Descriptive statistics | Population based on admission type |
|------------------------|------------------------------------|
|                        | All  | Direct | Prep  | CWP pilot population |
|                        | N    | 234   | 62    | 102   | 48                  |
|                        | Mean score (%) | 0.30 ± 0.01 | 0.38 ± 0.02 | 0.27 ± 0.01 | 0.28 ± 0.02 |
|                        | Min. score (%) | 0.03  | 0.10  | 0.03  | 0.10                 |
|                        | Max. score (%) | 0.77  | 0.77  | 0.63  | 0.60                 |
|                        | Variance (σ²) | 183   | 264   | 117   | 111                 |
| FCI Pre-test rank-sum Z-statistic | Direct | —     | 4.59  | 3.49  |
| IELTS Overall rank-sum Z-statistic | Direct | —     | 5.13  | 6.94  |

\( a \ p < 0.001 \\
\( b \ p < 0.001 \\
\( c \ p \sim 0.66 \\
\( d \) see Tab. III for a concordance with TOEFL IBT Scores

\( e \ p < 0.001 \\
\( f \ p < 0.001 \\
\( g \ p < 0.001 \\

TABLE III. Reproduction of Table 7 from Ref.39 giving a concordance of TOEFL IBT and IELTS test scores.

| IELTS Score | TOEFL Score |
|-------------|-------------|
| 9.0         | 118-120     |
| 8.5         | 115-117     |
| 8.0         | 110-114     |
| 7.5         | 102-109     |
| 7.0         | 94-101      |
| 6.5         | 79-93       |
| 6.0         | 60-78       |
| 5.5         | 46-59       |
| 5.0         | 35-45       |
| 4.5         | 22-34       |
| 0-4.0       | 0-31        |

D. Combining data from the KU Sharjah Campus

Beginning in Fall 2010, data of the kind presented above has also been gathered from the university’s campus in nearby Sharjah, UAE. We now ask of the combined data if there are any statistically significant differences in average pre-test scores. The above analysis for the larger AD campus data set gives good evidence that the respective campus populations should first be broken down by admissions type before being compared (e.g., direct entry students are compared to direct entry students across campuses, etc.). Also, we apply further statistical tests to determine if the AD data is normally (Gaussian) distributed, so as to avoid more computationally-intensive, non-parametric tests, like the rank-sum test. Table IV shows the results of tests to determine if the
TABLE IV. Normality and $F$-tests for AD campus pre-test data.

| Normality and $F$-test | Population based on admission type | All | Direct | Prep | CWP pilot population |
|------------------------|-----------------------------------|-----|--------|------|----------------------|
|                        |                                   | 263 | 183    | 1    | 107                  |
|                        |                                   | 17  | 35     | 107  | 4.34                 |
|                        |                                   | 0.6245 | 0.6682 | 0.6626 | 0.3113 |
| FCI Pre-test           | $K^2$-test*                       | 183.2 | 263.6  | 116.9 | 107.7 |
|                        | $K^2$-test*                       | 35.50 | 4.34   | 1.24  | 10.78               |
| F-test p-values        | Direct                            | —   | —      | 0.0001 | (0.0007)  |
|                        | Prep                              | 0.0001 | —      | —      | (0.7454) |
|                        | CWP                               | (0.0007) | (0.7454) | —      | —      |
| IELTS Overall*        | $K^2$-test                        | 17.11 | 0.67   | 23.44 | 1.04 |
| F-test p-values        | Direct                            | —   | —      | (0.9640) | 0.0078 |
|                        | Prep                              | (0.9640) | —      | —      | (0.0064) |
|                        | CWP                               | (0.0078) | (0.0064) | —      | —      |

* specifically, the D’Agostino-Pearson Omnibus $K^2$

AD campus pre-test data are normally (Gaussian) distributed and if so, how their variances compare. The first test applied is the D’Agostino-Pearson Omnibus $K^2$-test which is itself the combination of two tests, comparing both the skewness (asymmetry) and the kurtosis (peaked-ness) of the data as compared to the normal distribution. The null hypothesis in this case is that the data are indeed normally distributed. If true, the test statistic $K^2$ is chi-squared distributed with 2 degrees of freedom. Clearly, for FCI pre-test and IELTS data, the scores for all students together, regardless of population, are not normally distributed ($p < 0.001$ and $p < 0.005$, respectively). This is expected and consistent with our finding above, that the mean performance of the sub-populations is distinct and the differences between the mean scores are statistically significant.

When addressed separately, some assessment data for some of these sub-populations are consistent with normally distributed data. FCI pre-test scores for direct entry students and preparatory students not involved in CWP are consistent with being normally distributed ($p \sim 0.11$ and $p \sim 0.54$, respectively), however pre-tests for CWP students are likely not ($p < 0.005$). For IELTS scores, direct and CWP data are likely to be normally distributed ($p \sim 0.43$ and $p \sim 0.60$, respectively), but those of other preparatory students are not ($p < 0.001$). To this last point however, we notice in the raw data for preparatory students that there are some IELTS and FCI scores that are reasonable candidates for rejection as outliers.

We apply Chauvenet’s criterion to investigate possible outliers. IELTS scores for preparatory students show a ‘second bump’ at IELTS 7.5. Out of 105 such students, five have IELTS scores scattered above 7.0. To apply Chauvenet’s criterion for the set $\{x\}$ of $N$ data we calculate $n$ such that

$$n = N \times \text{Prob(outside } \pm \sigma),$$

where

$$t = \frac{x_{\text{suspect}} - \bar{x}}{\sigma},$$

$x_{\text{suspect}}$ is the suspicious data point, and $\sigma$ is the standard deviation assuming the data is normally distributed. The parameter $n = 2.7$, 0.4, and 0.14 for IELTS scores 7.5, 8.0, and 8.5 respectively. Whenever $n < 0.5$, there is reason to consider rejecting a data point from a set that one expects should be normally distributed. If we then reject both the 8.0 and 8.5 IELTS scores, the $K^2$ value for the preparatory IELTS data calculated in Tab. IV drops dramatically, from 23.44 to 15.87. If we further reject two of the IELTS 7.5 data points, consistent with Chauvenet’s criterion, the $K^2$ value drops negligibly. If we drop all five data points with IELTS 7.5, the $K^2$ value falls to 4.14 (with corresponding $t \sim 0.12$). So, while rejecting these seven data is not strictly justifiable, doing so does show that some of the conclusions as represented in Tab. IV blindly calculated, are weaker than they may appear.

We observe a larger sensitivity for the $K^2$ statistic for FCI pre-test scores for CWP pilot students. One student scored $3.2\sigma$ above the mean score (26%). Applying Chauvenet’s criterion, we get for the parameter of merit $n = 0.10$, so the data point meets the rejection criteria. Doing so and recalculating the $K^2$ statistic produces a large change, from $K^2 = 10.78$ to $K^2 = 4.55$, which has a corresponding p-value of $p \sim 0.1027$. Therefore, the
null hypothesis, that the data are normally distributed, can no longer be confidently rejected.

Therefore, given the overall small population sizes and the presence of arguable outliers which can dramatically weaken evidence of non-normality (expected in such an exotic student population), we hereafter assume that FCI and IELTS data are normally distributed to a reasonable approximation. Furthermore, when comparing and combining Abu Dhabi and Sharjah campus data, we feel somewhat justified in using the simpler Welch’s t-test, and not the more computationally taxing rank-sum, for comparing mean scores.

Now, we return to our original question: are there any statistically significant differences between student scores on the Abu Dhabi and Sharjah campuses? Table \( V \) shows FCI pre-test, post-test, and Hake’s gain scores for two course offerings where the FCI was administered on both Abu Dhabi and Sharjah campuses. Anticipating the possible existence of a gender gap, we also compared scores for both genders. For pre-test and post-test data, all scores gathered are included, but for gain scores, strict matching is applied which is why population sizes for some gains is smaller than either pre- or post-test sizes.

Symbols in the column headings, \( N \), \( S \), \( \sigma \), and \( \Delta \), stand respectively for the population size, the average test score for that row, the standard deviation in those scores, and the difference between the average Abu Dhabi and average Sharjah campus score. The last column shows \( p \)-values obtained as the result of a Welch’s t-test, where the null hypothesis is that the mean test scores on the two campuses are equal. Clearly, there is no significant evidence that student FCI scores and gains differ for the two campuses. When the population is broken down into the categories identified in the above analysis (preparatory and direct entry), remaining differences lose all statistical significance, however the sample sizes are limited, so it is difficult to attribute cause. So, all that can be said at present is that there are little to no statistically significant differences between Abu Dhabi and Sharjah campuses, in terms of students or treatments (instruction), as measured by FCI. Consequently, we hereafter merge both Fall 2010 data sets from the two campuses and both Spring 2011 data sets from the two campuses. We make a detailed comparison between traditional instruction and our CWP treatment. When CWP scores and gains are mentioned below, we mean those from the combined Abu Dhabi and Sharjah data sets for Spring 2011.

### E. Demographic Profile of Students

To briefly summarize this section, we see compelling evidence that the university population is unusually heterogeneous. There are substantial and statistically significant differences between two major sub-groups which are largely captured by categorizing them in terms of their admissions status: (1) direct admission, or (2) conditionally admitted via the preparatory program. Adding demographic data to the above analysis of test scores provides further insights. Table \( VI \) gives a comparison of some of the salient demographic features of U.S. engineering students and KU students. Combining this data with that of the analysis above shows clearly that the KU student population is unique, certainly different from the typical population for which interactive engagement techniques were originally developed (i.e. students of freshman, calculus-based physics, mostly engineering or pre-engineering majors in the US).

### IV. Designing the Collaboration Workshop

In this section, the important features of the initial internal proposal and the design process that was followed when creating the course are reviewed. At the time of that proposal’s writing (summer of 2010), our group at KU elected to take an approach that approximated engineering design. The overarching goal of the design was to provide ‘proof-of-principle’, to show that interactive engagement teaching could be adapted to a UAE context and that by changing pedagogy alone, greater student learning is possible. This began with crafting a problem statement using the then-existing baseline data presented in Sec. [III]. Objectives, constraints, and significant risks for any possible alternative pedagogy were then added to this problem statement, allowing for boundaries of the ‘solution space’ to be defined. Last, the group enumerated available assets, such as institutional resources, faculty and staff skills and experience, and student traits, in order to explore how each could be used in converging on a plausible pedagogical alternative that exploited those assets while remaining within the constraints. The concept and discussion of frames of context by Finkelstein and Pollock in Ref.2 proved a very helpful perspective for parsing diverse stakeholder values and maintaining focus only on those values associated with Type II and Type III transfers of norms and avoiding tempting Type IV transfers (see Sec. [II]). The cultural values and expectations shared by the majority of students, that have their origin in the larger cultural context of UAE society, would have to be incorporated and capitalized on in the design of a successful alternative pedagogy. The impetus for making this approach work and to succeed in reforming the physics instruction is obvious from Tab. [VI]. Conditionally admitted students, who are predominantly UAE nationals, make up 70% of the university student body and show alarmingly little response to traditional lecture instruction, to the extent that reforming pedagogy became a moral imperative.

What follows below are student learning objectives (L), boundary constraints (B), cultural objectives (C), and risks (R) identified by the our group. Each are listed with a brief discussion of a values/expectations-based justification. A reformed instructional strategy with reason-
TABLE VI. Comparison of demographic and learning gains data from traditional, first-semester, calculus-based physics courses for U.S. and KU engineering students.

| Semester       | Test     | Population | N^AD | N^Shj | S^AD (σ) | S^Shj (σ) | | \[p\]-value^a |
|----------------|----------|------------|------|-------|----------|----------| | |
| FCI Pre-test   | All      | All        | 68   | 29    | 0.29 (0.12) | 0.23 (0.11) | 0.06 | 0.0311 |
|                | Direct   | 23         | 6    | 0.36 (0.15) | 0.16 (0.06) | 0.20 | 0.1191 |
|                | Prep     | 40         | 23   | 0.27 (0.08) | 0.25 (0.11) | 0.02 | 0.4781 |
|                | Males    | 34         | 25   | 0.31 (0.14) | 0.23 (0.12) | 0.09 | 0.0474 |
|                | Females  | 34         | 4    | 0.27 (0.10) | 0.25 (0.04) | 0.02 | 0.5472 |
| Fall 2011 FCI Postest | All | All        | 64   | 28    | 0.36 (0.21) | 0.28 (0.09) | 0.08 | 0.0354 |
|                | Direct   | 22         | 6    | 0.57 (0.22) | 0.24 (0.08) | 0.33 | 0.1053 |
|                | Prep     | 40         | 22   | 0.26 (0.09) | 0.30 (0.10) | 0.04 | 0.1905 |
|                | Males    | 31         | 24   | 0.39 (0.23) | 0.29 (0.10) | 0.09 | 0.2968 |
|                | Females  | 33         | 4    | 0.34 (0.19) | 0.23 (0.06) | 0.12 | 0.2216 |
| Spring 2011 FCI Postest | All | All        | 65   | 21    | 0.35 (0.17) | 0.45 (0.21) | 0.09 | 0.3508 |
|                | Direct   | 0          | 0    | —     | —         | —        | | |
|                | Prep     | 48         | 21   | 0.28 (0.11) | 0.36 (0.22) | 0.08 | 0.3612 |
|                | Males    | 33         | 12   | 0.26 (0.10) | 0.39 (0.25) | 0.13 | 0.3207 |
|                | Females  | 31         | 9    | 0.29 (0.11) | 0.31 (0.18) | 0.03 | 0.7591 |

\[^a \text{In each case, the null hypothesis associated with the listed p-value is that the average FCI scores are equal.}\]

TABLE VI. Comparison of demographic and learning gains data from traditional, first-semester, calculus-based physics courses for U.S. and KU engineering students.

| U.S. Engineering Students | KU Direct Admission | KU Conditional Admission |
|---------------------------|---------------------|-------------------------|
| 19.5\% women              | 70\% women          | 35\% women              |
| 82\% Caucasian or Asian   | 80\% other MENA^a   | 90\% Gulf Arab (UAE)    |
| 13\% under-represented minorities | 20\% African or Asian | 10\% African or Asian |
| 1-in-5 are 1st generation comparable to U.S. | 1-in-3 are 1st generation |
| in their family to attend college in their family to attend high school |
| 1-in-9 not a U.S. citizen  | 9-in-10 not UAE citizens | 1-in-10 not UAE citizens |
| 17\% are ELL students >90\% are ELL students 100\% are ELL students |
| 8.0 IELTS is 'native' speaker 6.5 average IELTS 5.7 average IELTS |
| \(g\)_{Hake} \sim 0.22 \pm 0.02^d | \(g\)_{Hake} = 0.20 \pm 0.05 | \(g\)_{Hake} = 0.03 \pm 0.03 |
| DFW rate: 10 – 20\%^d | DFW rate: 7\% | DFW rate: 50\% |

\[^a \text{Middle East and North Africa}\]

A. Student Learning Objectives

1. *L1 Deliver the same content:* The new pedagogical approach should deliver the same topical/subject matter content that was presented by the traditional lecture version of the course. The reasons were two-fold: (1) to avoid the need to amend the official course syllabus with either the University-wide curriculum committee or the UAE Ministry of Higher Education and thereby delaying or jeopardizing the reform project, and (2) to limit the reform project’s exposure to the criticism that substantial improvements in learning gains were achieved solely by covering less material and focusing on less concepts for longer periods of time.
2. **L2 Reduce the course failure rate by half:** To demonstrate value-added to the university, the new pedagogy should make a significant positive impact on the student retention, by lowering the then course DFW rate from ~40% (mostly accounted for by UAE national students) to 20% or less.

3. **L3 Increase student problem-solving ability on ‘end-of-chapter’ style problems:** The new pedagogical approach should increase the ability of students to solve ‘end-of-chapter’ type problems by 10%, in order to bring exam performance up to a level consistent with achieving L2. The reasons for this requirement were two: (1) Universities in the UAE exercise a much larger level of oversight and quality control on summative examinations, consistent with their British-style inspirations, such as with exam writing committees, external examiners (3rd party graders) and periodic course file reviews by the UAE Ministry of Higher Education. In most of these contexts, exam writing norms consistent with the traditional lecture method (though not necessarily consistent with British-styled curricula) are enshrined in policy. Therefore, it would be difficult for instructors to use exams with a large fraction of purely conceptual assessment tasks. And (2), the group wanted to similarly limit the reform project’s exposure to criticism that increased problem solving performance was achieved by giving students problems that are perceived by many in the community as “easy, conceptual questions”.

4. **L4 Provide hands-on laboratory experience:** Most of our conditionally admitted students have had no exposure to experimental science in their secondary school curriculum. Furthermore, a well-designed kinesthetic experience is often most useful for addressing common misconceptions in introductory mechanics. Therefore, it is deemed critically important to have strong laboratory components to instruction vertically integrated through the entire college. First-year chemistry and physics courses must provide an effective cornerstone for students’ development of laboratory skills.

5. **L5 Double conceptual learning gains:** To line up with the metrics (in this case FCI learning gains) that were being used to form the basis of the proposal’s problem statement and the motivation for the reform project, the new pedagogy must substantially increase conceptual learning gains.

6. **L6 Increase student performance in core engineering science courses:** All of the teaching done by the Department of Applied Mathematics and Sciences is service teaching to the College of Engineering, making the engineering departments primary stakeholders. Therefore, for the reform project to gain long-term credibility and sustainability, it should demonstrate that it produces students who are better equipped at navigating the engineering curriculum, in courses such as engineering statics, engineering dynamics, circuits, etc.

### B. Boundary Constraints

1. **B1 Not change the contact time model:** As discussed in Sec. III the contact model for the traditional lecture-centered version of the course was “3+1” inclusive, meaning their are 3 credit hours available for lectures and 1 credit hour (3 contact hours) available for a laboratory and these two components are a single, 4 credit hour course. This contact model, which is enshrined in policy, is another example of an institutional norm that implicitly supports traditional lecture-centered instruction. This is evident from the fact that most reform projects at other universities have changed the contact time model for pedagogical reasons (e.g. Ref.4).

2. **B2 Not require new teaching staff:** Due to the start-up nature of the university and the recruiting challenges that established institutions in the UAE face, times scales for identification of new staff needs and initiation of hiring searches are typically twice as long as US universities. Therefore, any new pedagogical approach could not require additional teaching staff without facing significant delays.

3. **B3 Not purchase new laboratory or teaching technology:** Again, due to the start-up nature of the university and long turn-around times for procurement, purchasing and customs clearance in the UAE, any new pedagogical approach would have to function with existing laboratory and classroom equipment. The reform project was funded, but spending that funding and adding value through equipment purchases proved very difficult.

### C. Cultural Objectives

1. **C1 Use group work based on effective single-gender teams:** It was decided early in the course design process that mixed gender teams, while desirable in nearly every team-based, PER-based pedagogy, were an unnecessary risk in the KU context. Having teams of all male and teams of all female students working in the same classroom space was considered to be an appropriate middle-step in a spectrum of gradual gender integration across the students’ university career. In high school, for UAE nationals, the genders are geographically separated and never see each other. In preparatory program instruction, they learn passively and alone in mostly traditional instructional settings, but with the opposite gender present in the same room. And
in upper-class courses, as engineering majors, they all work on capstone design projects in mixed gender teams. Furthermore, as shown in Sec. III, there is no evidence of a significant gender-gap in FCI performance (with a larger sample size, it is likely that female students slightly out-perform their male peers) but there is quite a large performance gap between direct entry (mostly expatriate Arabs) and prep-track (mostly UAE national) students. Student-to-student interactions are essential for all IE pedagogies and the underlying hypothesis of teaming strategies, for those pedagogies that most heavily rely on extended group work, is that a diversity of ability levels and perspectives in the team members produces greater learning. In the KU context, if we are to assign students to teams and risk causing negative expectancy violation (since most students will want to work with family or tribe members, or close friends) it would be more beneficial for their learning to ask them to work with same-gender members of other families, tribal or ethnic groups, than to ask them to work with the opposite gender.

2. **C2 Used mixed ethnicity teams:** Conversely, as mentioned above, the most substantial performance gap on FCI and course exams is across the different admissions categories (and consequently, across ethnic backgrounds), not across genders. Negative sentiments should be minimized by making the purpose of the teaming as transparent as possible. Students must be made aware, both at the beginning of the course through an orientation, and repeatedly throughout the course, of the teaming recipe used and why it is important for their learning.

3. **C3 Conduct and document frequent self-assessments:** As mentioned in Sec. II and as suggested above by our objective to team students with other family/tribal groups, the course design must include regular feedback and performance monitoring, in anticipation of KU student’s demonstrated behaviors for special requests and complaints. Anonymous feedback surveys were seen by all as a way to gauge student discomfort levels early in the course and in a manner that would encourage them to express themselves more freely than they might in a face-to-face conversation. Also, continuous feedback and performance reports would allow the teaching team to respond using data to any complaints passed from students through university administrators.

### D. Risks

1. **R1 Avoid high faculty skepticism:** In the year that this project was proposed, the university held a series of faculty and staff workshops centered around the theme of innovative education in engineering. Outside presenters at the workshops consistently criticized lecture-centered instruction and showcased innovative approaches being developed and used at their home institutions in the US, Europe or the Far East. Anecdotally, there was clear evidence from questions and discussions at these workshops, that most faculty were very skeptical of the effectiveness of departures from traditional lecture-centered instruction. The skepticism took on two distinct forms or schools-of-thought. Those ascribing to the first generally considered the alternative pedagogy presented to be interesting, innovative and effective, but only for those students. These faculty were very skeptical, and for good reasons, that any such approaches would work for our students without major adaptation. Those of the second group hold an opinion more consistent with what physics education reforms in the West have addressed, namely, the belief that there is nothing fundamentally wrong with the lecture-centered approach itself (e.g. Ref. 45).

2. **R2 Avoid high student anxiety:** Interactive engagement approaches are known for raising student anxiety levels. In the context of an individual task, the instructors’ expectation of sense-making rather than answer-making calls for more abstract thinking, strains working memory more, and, in the process of queuing and confronting preconceptions, produces cognitive dissonance and mental discomfort more often than passively attending lectures or following a lab recipe. For KU students in particular, this is compounded by additional stressors at the course level, since students have rarely if ever experienced a course that is similar to interactive engagement (as mentioned above, many students have not even experienced a traditional recipe lab in their primary and secondary school science courses) and are engaging course content in a second-language. The risks of high student anxiety are serious. Students at KU have much more access and can provide one-on-one direct feedback to university administration about any issue. This can place significant administrative pressure on any course using an alternative pedagogy.

3. **R3 Avoid loose integration of interactive engagement activities:** Low learning gain IE courses reported in the literature usually report in the post-analysis that IE activities were used in a way that did not inform the broader norms of the course (see e.g. Refs. 46, 47). Feedback to students for heads-on/hands-on activities were ill-timed or feedback loops weren’t completed at all and/or conceptual tasks were not represented on high-stakes exams or were given little weight in the overall assessment. This gives students the impression that IE activities were integrated ad-hoc, did not re-
flect what the instructors ‘really’ expect of students, and were not a serious part of the course. In other words, the students believe the traditional instruction ‘is the real instruction’, and traditional learning (i.e. rote learning) ‘is the real learning’ that is expected, in spite of what the instructors are saying or doing. This risk was and remains very significant in the UAE context partly due to the exam writing/reviewing idio-culture mentioned above (see L3, Sec. IV A).

4. R4 Avoid lack of IE teaching experience: None of these authors have advanced degrees in PER. Two of these authors (GWH and AFI) have some experience (see Refs. 18 and 19, respectively) using PER-based instructional strategies.

E. Course design evaluation

Nine different interactive engagement approaches were considered for possible full, partial, or hybrid implementation. These were:

1. Cooperative Group Problem Solving (CGPS)6–10
2. Just-in-time Teaching (JiTT)50
3. Modeling Instruction (MI)51
4. Peer Instruction (PI)52
5. Student-Centered Active Learning Environment for Undergraduate Programs (SCALE-UP)53
6. Socratic Dialog Inducing Laboratories (SDI Lab)54
7. Tutorials (Ttls)55–57
8. Workshop Physics (WP)58,59

Table VII shows the results of our group’s considerations for eight different published IE instructional strategies. Each of the goals listed above and the means to achieve that goal are given a row in the table. Across columns, references are given, where they could be found, of evidence for (√) or against (X) that approach supplying the means specified by the desired goal and function. While entries are based on the references shown, the relevance and interpretation of the evidence presented is influenced by the subjective judgment of these authors and some of the decisions listed deserve some discussion.

1. Cooperative Group Problem Solving

As shown in Tab. VII the CGPS approach leads in positive evidence that it will match the needs of the KU context and scores relatively low on counts of negative and no evidence. Regarding positive evidence, CGPS requires no course content coverage reduction (L1), includes its own lab curriculum (L4), is adaptable to a “3+1” contact time model (B1), is given the designation of being a “low-effort” methodology (B2), and features a ‘traditional-looking’ course structure all because it was created so by design. CGPS was created by starting with the traditional course structure and iteratively changing the function of the individual parts within that structure. As summarized by Heller & Heller, “…we have been developing a conservative model that conforms with the usual structure and focus of the large introductory physics course…” and “The Minnesota model is based on the familiar triad of lectures, laboratories, and recitation (discussion) section” (Ref.19, p.4). The two instances of negative evidence for CGPS (B3 and R3) are due to, respectively: (1) the heavy use of video cameras, equipment that KU does not have, in their problem-solving labs and (2) the design process followed by CGPS; starting with the traditional course structure, then modifying it iteratively. It’s well-known that in the traditional, large-lecture physics course, a topic is treated in lab weeks apart from the lecture, and it was felt by these authors that this means “tight integration” (R3) is not accomplished by design if the traditional contact time model is followed. This does not mean lab and lecture are not synchronized in practice, just that it is not an automatic consequence of the course design. Perhaps most important, in terms of cultural expectations at KU, the CGPS method has already been studied with gender issues in teaming (C1) and efficacy with under-represented and at-risk student groups (C2) in mind (e.g. Refs.8,61 respectively), and has evidenced positive improvements in retention and learning.

2. Just-in-Time Teaching and Peer Instruction

JiTT, and similarly Peer Instruction, have many positive features that we anticipate would be quite beneficial if implemented in the KU context. However, there are a few facts that removed them from consideration. The foremost reason is that a learning management system (LMS; e.g. WebCT, Blackboard, Moodle, etc.) is essentially a prerequisite for implementing JiTT or PI, as the LMS is used to administer pre-class reading comprehension quizzes. At the time of our reformed course design exercise, KU did not have an LMS which is indicated as negative evidence for goal B3. Furthermore, a JiTT or PI implementation would still require adoption or creation of a separate laboratory curriculum (L4). For PI in particular, there is also often the need to reduce course content coverage (L1) to give time to in-lecture peer discussions. KU has since implemented Moodle for its LMS, so it is very likely that JiTT, and specifically PI, will be considered in future efforts to more deeply reform the lecture portion of the course.
TABLE VII. Evaluation matrix for published IE pedagogies considered for the KU context.

| Goal | Means to achieve goal | Published evidence that a method provides needed means |
|------|-----------------------|-------------------------------------------------------|
|      |                       | CGPS | JFTT | MI | PI | SCALE-UP | SDI Labs | Tutorials | WP |
| L1   | Requires no coverage reduction | ✓    | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| L2   | Efficacy with URM & 'at-risk' students | ✓   | ?   | ?  | ?  | ?   | ?       | ?        | ?      | ?  |
| L3   | Increases trad. problem solving skill | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ?      | ?  |
| L4   | Includes effective lab curriculum | ✓   | ⃝  | ✓  | ✓  | ✓   | ✗       | ✗        | ✗      | ✗  |
| L5   | Increases FCI (or FMCE) gains  | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✗      | ✗  |
| L6   | Attract/retain more engineering majors | ?   | ?   | ?  | ?  | ?   | ?       | ?        | ?      | ?  |
| C1   | Single-gender teams effective | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| C2   | Mixed-ethnicity teams effective | ✓   | ✓   | ?  | ?  | ?   | ?       | ?        | ?      | ?  |
| C3   | Allows freq. student feedback  | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| B1   | Adapted to “3+1” contact time model | ✓   | ✓   | ?  | ?  | ?   | ?       | ?        | ?      | ?  |
| B2   | “low-effort” methodology | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✗      | ✗  |
| B3   | No additional lab equipment needs | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| R1   | Can leave lecture unchanged  | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| R2   | Have 'traditional-looking' course elements | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| R3   | Elements tightly integrated by design | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |
| R4   | These authors have some experience | ✓   | ✓   | ✓  | ✓  | ✓   | ✓       | ✓        | ✓      | ✓  |

Totals:
- Counts of positive evidence (✓): 13 8 9 7 9 8 8 5
- Counts of negative evidence (✗): 2 2 4 5 6 5 1 6
- Counts of no evidence (?): 1 6 2 4 1 3 7 5

3. Modeling Instruction, SCALE-UP, SDI Labs, and Workshop Physics

Modeling Instruction, SCALE-UP, SDI Labs and Workshop Physics are certainly the most experimentally-focused and kinesthetically engaging of those PER-based instructional strategies considered here. The performance of most of them has also been extensively studied in secondary implementations across high schools, colleges, and universities in the US. Consequently, these instructional strategies present some of the most attractive options available in the PER literature for achieving the learning goals identified for the KU context (L1-L6). Unfortunately, the radical changes to course structure that they respectively require are too institutionally risky at KU (B1-B3) to be attempted at present. Increased credibility created by “early wins” from a less taxing first-reform project may change this situation in the future.

4. Tutorials

Tutorials (Ttls) packages developed by University of Washington (UW), University of Maryland (UM), and Open Source (OS) versions were considered. The lack of a dedicated recitation hour in the existing KU course contact model meant any Ttls set would have to be implemented in a non-standard way which has been done previously in all three cases (hence satisfaction of B1). However, this is most often done by using them as group activities during lecture instruction which puts satisfaction of R1 in question. So, while Ttls are not considered a stand-alone solution to the course reform design problem, as a library of PER-based tasks they are very attractive for use. One author (AFI) has experience with UW Ttls and there is more data on their performance, so while Ttls in general were selected for hybridization into the CWP session (see Sec. [IV] below), UW versions were used (and modified, mostly to simplify language use) exclusively.

F. Collaborative Workshop Physics (CWP):
Course Structure Prototype

Following the above considerations, we converged on creation of a hybrid approach which we call ‘Collaborative Workshop Physics’ (CWP). The most important
constraint was that the lecture remain unchanged and so we focused on creating a reformed learning experience with the 3-hour lab period available each week. The principle components of the instructional strategy are taken from CGPS and various Tutorials however both were substantially modified to fit the context and design constraints. Aside from showing evidence that CGPS could supply the means needed for the course reform, context-rich problems (the central feature of CGPS recitation instruction) were seen very favorably because of their similarities with problems posed in innovative engineering design education. Specifically, design problems in engineering that are considered salient are often described in engineering education literature as being “realistic”, “ill-posed”, “ill-structured” or “open-ended” which are terms also used to describe context-rich problems in physics. Atypical of common solution strategies for traditional end-of-chapter or analysis problems, both kinds of problems also require skills like; tolerance for uncertainty, estimation, big-picture thinking, self-questioning for clarification, teamwork, and multiple representation use (see Ref.10, Sec. 11 “On Design Thinking” for example). It is reasonable to assume that both kinds of problems call upon similar cognitive resources and produce similar cognitive loads, though the goals for each kind of problem are different (creating a product vs. creating a prediction). This similarity is attractive for physics course design for creating a “knock-on” effect that could indirectly benefit students in later engineering design courses and thereby positively contribute toward learning objective L6.

The next most important constraint, and one of the few reasons all CGPS features present in the University of Minnesota model could not be implemented “off-the-shelf” was the lack of necessary lab equipment and the inability to retool (B3). However, this is easily mitigated, since the CGPS recitation sessions and CGPS laboratory sessions were originally designed to operate independent of each other10. Therefore, we chose to create an instructional strategy that borrowed only recitation/problem-solving techniques from the CGPS model. That meant however, that goal L4, and possibly L5, could not be satisfied because there is no provision for a lab curriculum and limited opportunities in a CGPS recitation to queue mechanics misconceptions in a concrete, kinesthetic manner. To mitigate this, we took inspiration from the “box-of-probes” philosophy of Workshop Physics and the equipment already available on-site. From this perspective, the simplest way to replace existing recipe labs with a reformed version was to narrow the goals of the experiments and remove their given procedures.

We called these shortened labs experiment problems. They are given to the student teams in the form of a single, simple question and the main tasks are to reach a consensus on a measurement protocol, execute with the available measuring tools, and answer the question within 1-page only, through evidence-based reasoning with their measurement results and error analysis. The particular phenomenon investigated is chosen such that it is a concrete experience of a simple system sharing the same underlying physical principles involved in the context-rich group problem. By posing the experiment problem before the context-rich problem, students’ mechanical misconceptions are afforded an opportunity for concrete queuing, similar to the manner in which many such misconceptions are formed (kinesthetic experience during childhood sensory-motor development). Instruc-
tor coaching to teams, during their procedure design and when later solving the context-rich group problem, can then take on the form of short ‘Socratic dialogues’. To draw attention to, illicit reflection on, and provide targeted intervention against misconceptions. But it’s ultimately left to peer discussions, enriched by such periodic coaching visits, to discover the correct interpretation or solution, and proceed once consensus is reached.

Despite the intent on lowering the ‘learning curve’ for context-rich problems, the experiment problems alone were not deemed sufficient for effective preparation. It was decided to conduct a ‘warm-up’, that featured heavily scaffolded training with drawing representations, to proceed the experiment problem. The representation featured in the ‘warm-up’ exercise would again be the same one most useful for thinking through the experiment problem and the context-rich problem. However, the exercise would be ‘drill-like’, ungraded, and used primarily to teach the student how to clearly draw that representation and establish correspondence between its features and the features of an example physical system.

Figure 6 shows a flow chart representation of the Collaboration Workshop instructional strategy. Prior to the first session, students meet in their lab sections twice; once to take a small battery of pre-instruction assessments (FCI and a mathematics test) and the second time to receive individual training with some basic tools (meter stick, vernier caliper, digital stop watch, digital scale, photogate/light barrier with digital signal box, air track with glider carts). The following core features of the CGPS instructional strategy are followed thereafter, but with the following caveats:

1. Formation of well-functioning cooperative groups. Students are assigned to teams matching the ideal composition reported in Ref. and conforming to context constraints (C1, C2). Each team is gender-homogeneous and composed of 4 students having a heterogeneous skill distribution (based on FCI pre-test, a pre-calculus math test, and IELTS score drawn from admissions records). Each team contains at least one member having a relatively high FCI pre-test score, one having a relatively high math pre-test score, one having a high IELTS score, and one who is an expatriate. Each member is given a rotating assignment to one of four different roles: Manager, Recorder, Skeptic, or Idea Generator. The role of Idea Generator is modified from that of Explainer in that they are coached to think of as many different solution strategies as they can for their team’s problem, in addition to clarifying. As mentioned above, this is motivated by the desire to model similar habits of mind also found in effective engineering design education, where solving design problems is understood as a “divergent-convergent” questioning process. Idea Generators are coached to personally introduce, and to illicit from others, as many solutions to the team’s problem as possible (i.e. divergent thinking). Students in the remaining roles are coached in the standard way. Skeptics are coached to ask probing questions in search of weaknesses in the physics behind ideas, with the goal of convincing the team to eliminate them from consideration (i.e. convergent thinking). Managers are coached to facilitate team dynamics; keeping the team on task, organizing work into sub-tasks, and sequencing the team’s work on the task in a logical order. Recorders are coached to check each member for their agreement on a plan of action and to only record what is reached by consensus. Roles are rotated after four weeks (four sessions) and students are reassigned to new teams based on available formative assessment data (CWP session average, in-class quiz and exam scores) in addition to pre-test assessment and language level.

2. Repeated reinforcement to use a prescribed problem-solving strategy. Students are explicitly taught a problem-solving strategy that is essentially the same as that described in Ref., though with a few minor innovations. The sequence taught is “read - imagine - draw - graph/diagram - chose theory - measure/calculate” and it is reinforced with each of the three components of the CWP sessions (tutorial, experiment problem, context-rich problem). Reading is explicitly added and coaching attention is specifically devoted to it in the strategy since the nearly all KU students are ELLs (see Tab. and often miss important information that is not represented numerically in the problem statement. The effect of this is not unlike that of other well-known student attitudes about problem-solving (e.g.), but the combination of both novice attitudes toward problem solving and the second-language learning environment mean ‘plug-n-chug’ solution strategies are very common and very difficult to correct without persistent intervention. Lectures for the course, while largely unchanged in terms of content and format (goals & means LI, B1, and R1), are paced so that new concepts are introduced by the CWP session and reviewed and abstracted upon by the lecture. Example problems demonstrated during lecture model the same problem solving approach.

3. Using ‘context-rich’ problems as opposed to ‘end-of-chapter’ problems. This feature of the CGPS instruction is implemented unchanged and some of the problems in the CGPS manual were used, though sometimes modified for context (i.e. changing regional terms like “state trooper” to the more generic “police officer”, etc.). Other problems used were designed following the recipe in Ref.

4. Individual accountability via positive interdependence via goal interdependence. This feature of the CGPS instructional strategy could not be implemented as recommended in Ref. There, the
For two gliders before the collision in the figure below, assume that $m_A = 2m_B$ and compare the magnitude and direction of the following quantities:

- the net forces on the two gliders at an instant during the collision
- the changes in momentum of each of the gliders
- Also find the kinetic energy of the gliders before and after the collision

\[
\begin{align*}
m_A v_A & \quad m_B v_B \\
\end{align*}
\]

**ACTIVITY 2. EXPERIMENT PROBLEMS (90 min)**

By placing two glider carts on the air track, we can study collisions between them. As before, light gates allow us to determine their velocities before and after they collide. Depending on the material at the point of contact (metal, rubber, clay, springs, magnets, etc.), the gliders can exert a wide variety of forces on each other.

**Lab question:** Is there anything that does NOT change, no matter what material, no matter what force we choose?

**Lab problem:** What is the change in the momentum of the two-glider system for a variety of forces present during their collision (each group should pick just one force, not the same as your neighbor)?

**ACTIVITY 3. CONTEXT-RICH PROBLEM (60 min)**

As an employee of SEMA (Save Earth Mission Agency), you are analyzing a collision between a space probe and an asteroid. Data log from the space probe tells you that the probe was moving with the speed $v_p$ right before it collided with the slowly moving asteroid at the far edge of the Solar system. Estimating that the mass of the asteroid is about three times the mass of your space probe. You guess that something must have gone wrong with a guidance computer on board, but you also need to check which of the several possible SEMA-suggested scenarios for the dynamics of the space probe and the asteroid after the collision fits the situation most closely?

\[
\begin{align*}
&i) \ v_{x1} = 0, \quad v_{x2} = 1/2 v_p \\
&ii) \ v_{x1} = 1/4 v_0, \quad v_{x2} = 1/4 v_p \\
&iii) \ v_{x1} = -v_0, \quad v_{x2} = 2/3 v_p \\
&iv) \ v_{x1} = -1/2 v_0, \quad v_{x2} = 1/2 v_p \\
\end{align*}
\]

FIG. 7. An example sequence of activities in a CWP session, in this case, for instruction on linear momentum.

The authors compared two different strategies they tested for creating positive interdependence within groups (goal interdependence and reward interdependence), with the objective of fostering mutual concern for individuals’ success within groups and personal accountability to contribute toward the group effort. They found that reward interdependence, created by adding a group problem solved in recitation to the score of individual in-class exams, was superior in this regard. For logistical reasons and reasons explained in Sec. IVA for goal L3, we were unable to implement this and instead relied on goal interdependence, by requiring groups to reach consensus on major outcomes and report only the consensus on session deliverables.

G. Collaborative Workshop Physics: Example Sequence of Session Activities

Figure 7 illustrates an example set of activities used in the CWP sessions and helps to explain how they are chosen to form a coherent sequence. In this example, students are working through a variety of tasks revolving around linear momentum and conditions for its conservation. First, notice that in all the tasks, there is little or no numerical information given. This is done to reinforce explicit addition and attention to reading in the problem solving strategy. Students are instructed to keep their pens and pencils down for the first 5 minutes of each activity and to read only. During this time, the instructors make their first round of visits, asking students to simply answer questions like, “What is the big idea?” and “What does the writer of this problem want from you?” “What is the goal your team needs to reach?” Feedback and coaching is given on text analysis. For example, many students struggle with the multiple uses of the word “momentum” and its apparent derivatives (i.e. “momentum”). Some often, quite reasonably, conclude that the word is used in reference to “time” (i.e. a very short time interval), rather than to torque or momentum. This first round of coaching allows instructors to engage in short discussions about context and how context in the problem statement can modify the meaning of jargonized words such as these. If necessary, the class will be stopped for a few words from the instructor if an issue appears to be common to all groups. One instructor recalls in their journal a 5 minute discussion of the word “hammer” for the case when it is used as a verb (i.e. ‘to hammer a nail into wood’) rather than as a noun to identify the tool.

For the tutorial, coaching attention is more individual and direct, and emphasis is placed on drawing graphical representations that effectively and compactly contain the information given in the text. Instructors visit teams a second time, just prior to finishing the tutorial, and this time ask questions about the sizes and directions of vectors, give small follow-up tasks, like drawing a corresponding kinematic graph, and encourage members to compare and discuss their results. Every student in the group is given a tutorial sheet to work with, but an additional copy is given to the group’s recorder to sketch a consensus drawing.

Following the tutorial, a single sheet with the printed experiment problem is given to the Recorder in each group. Experiment problems were created in such a way that neither complicated measurements or lengthy data analysis were necessary for a satisfactory solution. Rather, a simple, concrete question is posed about objects and/or motion presented/demonstrated to them, but with no procedure and no suggestions for specific equipment to be used for making measurements. Teams are required to create a simple procedure, justify it physically and to submit a concise, written (one-page, front-only), evidence-based (measurements must be used in their argument) solution. As the task is handed out, special attention is again given to reading. Instructors follow a common coaching scheme for the experiment problems which is as follows:

- On the first visit
  - Strong encouragement to read problem statement
  - Discuss, “What is being asked of us?”
– Generate a large number of ideas for possible solutions
– Withhold criticism of each other’s ideas
– Strong discouragement to touch any lab equipment or make measurements for first 20-30 minutes of the problem

• On a typical second visit (10 minutes after discussions begin)
  – Socratic questioning of the team, to gauge and to guide clear understanding of the problem statement
  – Strong encouragement to begin eliminating the weakest ideas for solutions
  – Strong encouragement to balance time spent converging on solutions versus building an apparatus and conducting measurements

On third and later visits, instructors focused on questioning the team about features of their chosen solution strategy (e.g. “Why did you choose this detector over another?”) and what would happen if they made modifications (e.g. “What if we change the location of this photogate, what will happen to your graphs?”) Early in the course, during the first 1-2 CWP sessions, there was also often a need to stop the whole class and present a few tips for effective team work, to warn groups against common pitfalls. The most common is the tendency to ‘fall in love with the first idea’ and converge (usually under duress from a perceived lack of time) too quickly on executing a sub-standard solution strategy. At the conclusion of the experiment problem, teams were often encouraged to take a 5-10 minute break. Upon return, one context-rich group problem is distributed to each Recorder. Again, teams are coached to not write anything for the first 10 minutes or so, but rather to read, discuss, and answer amongst themselves, “What is the goal, what is the writer of this problem asking us for?” Instructors made a visit after this initial period to illicit reflection.

V. RESULTS FROM THE CWP PILOT

The results of the secondary implementation of PER-based teaching innovations described above are presented below as they relate to the learning objectives declared during the course design. The most directly measurable objectives are L2: reduce course failure rate, L3: increase ‘end-of-chapter’ problem-solving ability, and L6: double conceptual learning gains.

A. Reduced Course Failure Rate

A coarse-grained but important measure of instructional success is the course failure rate. As discussed above and as shown in Tab. VI conditionally admitted students show an alarmingly high course failure rate (50%) under traditional instruction. One goal for the reformed course, featuring the CWP instructional strategy, was to reduce this rate by half. The pilot offering was delivered exclusively to conditionally admitted students and as a consequence, was an ideal experiment to determine the efficacy of the new teaching approach. For students attempting the course for the first time, the average fraction failing for all previous, traditionally taught offerings has been 0.24 ± 0.06 (standard error of the mean, average is taken over course offerings), but for conditionally admitted students the fraction failing has been 0.50 ± 0.10. Coincidentally, the failure rate for the reformed course is 24%, but since these are all conditionally admitted students, there is evidence that CWP instruction lowers the course failure rate by about half, from 50 to 24%.

B. Improved traditional problem-solving skill

Table VII shows exam scores from a selection of course offerings that have been analyzed to date. The major determinant for the course failure rate is student performance on summative course exams. One goal for the reformed course was to increase the students’ skill at solving traditional ‘end-of-chapter’ style problems by 10% (L3), given problems of equal difficulty. This goal was motivated by and necessary to achieve the desired reduction in course failure rate. Of course, it’s non-trivial to establish the relative difficulty of a group of course exams, each with different problems, so that the performance on them can be compared. Our approach to solve this problem was to query groups of professors and students unaffiliated with the CWP project and ask them to perform a categorization and ranking task with the individual problems. Problems from the mid-term examinations, for all offerings that have been analyzed to date. The major determinant for the course failure rate is student performance on summative course exams. One goal for the reformed course was to increase the students’ skill at solving traditional ‘end-of-chapter’ style problems by 10% (L3), given problems of equal difficulty. This goal was motivated by and necessary to achieve the desired reduction in course failure rate. Of course, it’s non-trivial to establish the relative difficulty of a group of course exams, each with different problems, so that the performance on them can be compared. Our approach to solve this problem was to query groups of professors and students unaffiliated with the CWP project and ask them to perform a categorization and ranking task with the individual problems. Problems from the mid-term examinations, for all offerings that have been analyzed to date.

| Course Offering | Mid-semester | Final |
|-----------------|--------------|-------|
|                 | N            | S ± s.e. | N    | S ± s.e. |
| Fall 2009 (T)   | 65           | 0.61 ± 0.03 | 61    | 0.55 ± 0.03 |
| Fall 2010 (T)   | 53           | 0.61 ± 0.02 | 46    | 0.57 ± 0.02 |
| Spring 2011 (CWP) | 50         | 0.77 ± 0.03 | 50    | 0.70 ± 0.02 |

\[ \Delta S_{2011-2009} \pm \text{s.e.} = +0.16 \pm 0.04 \]

\[ \Delta S_{2011-2010} \pm \text{s.e.} = +0.16 \pm 0.04 \]

\[ +0.13 \pm 0.03 \]
rank the problems in each group according to difficulty. This was repeated with problems from the final exams as well. Analysis of this data is still underway, but preliminary indications are that: 1) a problem from a CWP exam is present in every category introduced, and 2) a CWP problem is ranked as the most difficult in nearly all categories. This evidence supports the assertion that the mid-semester and final exams used to assess students in the reformed course are more difficult than the exams in previous offerings. Consequently, directly comparing raw CWP exam scores to scores from previous offerings should provide a valid and conservative lower bound for improved, traditional problem-solving performance. With these caveats, improvement in traditional problem-solving ability, as shown in Tab. [V B] is approximately 15% and likely more.

C. Improved conceptual learning gains

Averaging over all traditionally taught \((T)\) offerings of the course, the normalized gain (or Hake’s gain) of conditionally admitted students on the FCI prior to CWP instruction was \(\langle g \rangle_{T,Hake} = 0.03 \pm 0.03\). The error here is the standard error of the mean. Figure 5, panel c., shows a histogram representation of the corresponding pre-test (black) and post-test (white) FCI distributions. The normalized gain using CWP instruction, including both offerings on the Abu Dhabi and Sharjah campuses, was \(\langle g \rangle_{CWP,Hake} = 0.14 \pm 0.04\), a modest improvement. Nevertheless, Cohen’s effect size \(d = 0.67\) for the two distributions of normalized gains is

\[
d = \frac{\langle g \rangle_{CWP} - \langle g \rangle_{T}}{\sigma_{CWP+T}} = 0.67,
\]

where \(\sigma_{CWP+T}\) is the standard deviation of individual student gains of both data sets. The value \(d = 0.67\) is moderately high, evidencing that the benefit of CWP instruction to the average student is statistically significant, enough to outweigh the relatively small sample sizes. Of course, there is also ample room for further improvement.

VI. DISCUSSION

In this section, we discuss implications of the results presented above, in terms of the evidence they provide for answering the major research questions of this work (see Sec. I).

A. Is lecture-centered instruction in the US and in the UAE contexts equivalent, in terms of course structure, execution, and effect on student learning?

The case presented in this work from KU provides compelling evidence that traditional, lecture-centered instruction is similarly structured and produces similarly poor student learning in both contexts. There are important caveats however, that are worth mentioning. Directly admitted students to KU engineering programs, while having similar learning gains to those of traditionally taught engineering students in the US \((\langle g \rangle = 20 \pm 5\%\) vs. \(\langle g \rangle \sim 22 \pm 2\%\) respectively), share very few demographic traits in common (see Tab. [V I]). Unlike US engineering majors, KU direct admission students are mostly female, mostly expatriates, and mostly ELLs. Given the growing body of significant research connecting issues of equity in instructional practice to gender and ethnicity (e.g. 61,62,70,79,80), it is quite possible that similarities in learning gains are not the result of similar causes. For instance, KU conditionally admitted students have a male-to-female student ratio that is closer to that of US engineering students and the normalized gain for traditional instruction is even lower, consistent with zero in fact. But it is not yet clear if the gender distribution is strongly correlated with this difference. The difference in language level proficiency (as measured by IELTS) between the two groups is likely equal if not more important. Also, while there have been some recent studies on reformed physics teaching with ELLs, there are very few reports of situations where the language of instruction is also the language of the majority outside of the classroom (such as Ref. 13). This is also the case at KU, where the language of instruction is mostly used in isolation on campus and opportunities for authentic use and reinforcement off-campus are very limited.

All of these issues warrant further research and offer potential for improved pedagogical innovations for...
Physics education in the UAE and other similar contexts. For the case of language, since we used language level proficiency in our teaming algorithm, we can draw some preliminary conclusions on the efficacy of this practice. The largest reduction in course failure rate is for students with IELTS overall scores of 5.5, none of whom failed the course. This group makes up 15% of the students in the pilot offering of the new course and, based on data from traditional instruction, half of them would be expected to fail. There are not yet enough statistics to make strong conclusions, but the fact that none of these students failed suggests a positive contribution toward achieving L2: efficacy with ‘at-risk’ students, recalling that IELTS 5.5 is the average score of conditionally admitted students.

B. Does interactive-engagement instruction in the US and UAE contexts require similar measures for implementation and produce similar learning gains?

To answer this question, one must first determine if interactive-engagement teaching has been achieved. As discussed in Ref.8, faculty engaged in a new reform effort often think that they are teaching in a manner consistent with findings in the PER literature when in fact the classroom norms they establish suggest otherwise. If one considers high normalized gain on pre-/post- instruction conceptual inventories as the ‘signal’ of interactive-engagement success, then the present case is not clear-cut. Hake reports that on average, successfully implemented IE courses in introductory mechanics have \( \langle g \rangle = 0.48 \pm 0.14 \) while traditional courses have \( \langle g \rangle = 0.23 \pm 0.04 \). In the KU case, \( \langle g \rangle \) for conditionally admitted students improved from 0.03 ± 0.03 to 0.14 ± 0.04, but both could still be categorized as traditional based on using Hake’s results as the definition of IE versus traditional.

In spite of this, these authors assert that interactive-engagement teaching has been successfully implemented. Substantial care was taken to thoughtfully adapt the CGPS instructional strategy for recitations to the constraints present for the KU first-semester introductory physics course and without compromising core features, including classroom norms, that are critical for reproducing its published effectiveness. To support our argument, first we point out that improvements in gain of the size seen at KU are similar to other cases where only laboratory instruction was reformed and where no evidence exists to suggest that PER-based norms were absent from the lab setting. For example, in Ref.81, mean raw gain on FCI is similarly increased by about 9% as a result of reforming only the laboratory portion of the equivalent course. Similar results are reported by Hake82 for the cases of UL-RM95S-C and UL-RM95Sna-C \((\langle g \rangle = 0.25 \text{ and } \langle g \rangle = 0.26, \text{ respectively})\) where only the labs were reformed, improving \( \langle g \rangle \) over the traditional course \((\text{UL-94F-C, } \langle g \rangle = 0.18)\) by 0.07-0.08. (see Ref.82 for details).

We further support our argument by suggesting that for KU and similar institutions (having large ELL populations), \( \langle g \rangle \) averaged over the entire class may be too course-grained of a measure to adequately ‘detect’ interactive-engagement teaching and the increased student conceptual learning resulting from it. We hypothesize the reason \( \langle g \rangle \) is not a clear measure in our case is due to a strong connection between learning gains and proficiency in the language of instruction (English). We evidence this by extending on the above discussion of language. Figure 9 shows \( \langle g \rangle \) for FCI versus IELTS overall score. Here we see a strong, positive correlation between IELTS score and \( \langle g \rangle \), for students taught traditionally \((R = 0.88)\) and in CWP \((R = 0.82)\). But at very high language level \((\text{IELTS 7.0, about } 10\% \text{ of either population})\), the difference between traditional and CWP instruction is large \((0.16 \pm 0.10 \text{ vs. } 0.47 \pm 0.08, \text{ respectively})\) and significant \((p < 0.001 \text{ via rank-sum test})\). Considering that IELTS \( \geq 8.0 \) is considered ‘mother-tongue’ proficiency14, one might roughly assume learning for these students to be largely free from second-language-related factors. If so, then it appears that traditional instruction to KU students with increasing English proficiency converges well to Hake’s result \((\langle g \rangle = 0.16 \pm 0.10 \text{ and } 0.23 \pm 0.04, \text{ respectively})\) in the limit of ‘mother-tongue’-like proficiency. And the same convergence for CWP to Hake’s IE result \((\langle g \rangle = 0.47 \pm 0.08 \text{ and } 0.48 \pm 0.14 \text{ respectively})\) appears in the same limit. All other major factors of the two groups are accounted for; both groups are conditionally admitted and have the same language ability. Both groups contain roughly equal numbers of students from all major demographic constituencies: males, females, expatriates and UAE nationals. Both groups have consistent expectations associated with learning physics course content. Thus, this increase in learning gains at IELTS 7.0, having controlled or accounted for other major factors, is strong.
evidence that gains were predominantly caused by pedagogical reform and that interactive engagement teaching was successfully implemented.

C. Can failure risks for secondary implementations of PER-based instruction be predicted and mitigated, even in the presence of large cultural differences with the primary context?

An analysis of qualitative data on classroom norms in the UAE, gathered from both these authors and from the literature concerning surrounding institutions, shows that indeed, there are certainly traits of the idio-cultural context for UAE universities that differ in kind to those encountered in the US. The role the university plays in the larger UAE society, and how that informs the motivations of perspective and current students, can have a strong influence on a student’s response to PER-compatible classroom norms. This is largely determined by the broader ethno/national-cultural context and cannot be addressed by instruction, nor is it necessary to do so, given the manner in which the expectations of UAE national students are modified in narrower, classroom contexts. What we present is a proof-of-principle: one can generate from qualitative data on cultural values, classroom behaviors, and institutional factors a set of culturally- and contextually-informed design requirements for a successful PER-based course, evaluate the suitability of published PER-based pedagogies against these requirements, and converge on one that likely has the highest chances for success upon roll-out. Like any proof-of-principle, this work offers no guarantee that the same approach, tried under the same conditions would succeed however, it should silence suggestions that such an adaptation is impossible and doomed to failure. Furthermore, a design-based approach like the one presented in Sec. IV naturally and logically suggests ways in which to thoughtfully modify a PER-based pedagogy while respecting its core functions, in this case for CGPS.

VII. SUMMARY AND OUTLOOK

We find that PER-based instructional strategies, originally developed in North America and Europe, can and should be adapted for secondary implementation in physics classrooms in the UAE. The modifications to a published instructional strategy that are needed to make a secondary implementation successful are not trivial, but neither are they impossible. We have presented a thoughtful analysis of the cultural values of our students and how they differ from those of students for whom PER-based instruction was originally designed. From this, we have extrapolated what values our students bring to university with them as freshmen, how they are modified by classroom context, and have converged on requirements for a reformed course design, triangulating on these conclusions using our own observations of student behavior and those of other educators in neighboring institutions. From this, we have followed an engineering design-based evaluation of existing PER instructional strategies and identified CGPS as best exploiting our existing assets and providing the greatest number of necessary core functions for a reformed course in our context. The same evaluation also helped identify beneficial modifications of CGPS and additions from the Tutorials-based approach which lead us to focus our reform effort on the laboratory and to create the Collaboration Workshop. We believe that institutions in the Gulf region and in the developing world in general, which are considering course reforms using PER-based tasks and instructional strategies, could benefit significantly from using this approach as a reference.

The need for such pedagogical reforms at KU was based on students’ unsatisfactory course performance and conceptual learning, in courses that are essentially ‘traditional’, as defined in the PER literature and that feature lecturing to passive audiences, algorithmic problem exams and verification recipe labs. The poor conceptual learning seen in North American studies for traditional course formats is made worse when implemented for UAE national student populations, who are predominantly ELLs and who ascribe a very different societal role to university. A pilot set of two offerings of the Collaboration Workshop approach, one on each of KU’s two campuses, created to address these differences has produced significant improvements in course DFW rate, traditional problem-solving ability and conceptual learning gains.

The Collaboration Workshop pilot has also produced valuable data that raises questions for future research, relevant for institutions similar to KU. Regarding issues related to pedagogy, the possible strong modulation of conceptual learning gains by language proficiency in the language of instruction requires further investigation. Distributing the high language level students was meant to help overcome language barriers between students and instructors and enrich peer interactions. Student teams for CWP were designed, whenever possible, to have at least one high English-language-level member, but these students appear to be the sole recipients of improved conceptual learning gains and not their other team members. Why is this the case? Also, distributing the high language level students is often combinatorially frustrated by the inability to create mixed-gender teams. Thus, another important direction for future research is to find other, similarly-effective, language-related coaching and support methods, for increased conceptual learning gains to students at lower language levels.

Regarding issues related to assessment, despite giving students their choice of FCI in English or Arabic, there is not yet enough evidence to confirm that the test has high validity with KU students. Are the normalized gains on FCI of lower language proficiency students hindered primarily by less conceptual learning, due to
language barriers faced during instruction, or is it due to these students’ inability to comprehend the FCI questions clearly? Many pre-test scores are alarming close to random. Are most students guessing? Anecdotally, the recent and rapid modernization of UAE nationals’ society means that for many students Arabic and English literacy levels are often comparable and equally underdeveloped. In other words, for written communication, there may be no ‘mother-tongue’ amongst many of our students. How does this impact the validity of the FCI and other conceptual instruments? If the effect is large, can and how should such an assessment instrument be modified to recover its validity? The Collaboration Workshop has also raised the on-campus visibility of pedagogical alternatives to traditional lecture. Though these authors have also raised the on-campus visibility of pedagogical alternatives to traditional lecture. Though these authors present this work as a proof-of-principle, do other faculty and administrators view it as such? What are ways to extend upon ‘early wins’ produced in the pilot that will allow for reforms to the lecture sections and course exams, which would presumably further improve student course performance and conceptual learning?

ACKNOWLEDGMENTS

The authors would like to thank Dr.s Moh’d Rezq and Jamal Hassan, for their help gathering FCI and MPEX data in their classrooms, and for their support and feedback during the CWP pilot implementation. We thank Dr.s Taufiquar R. Khan, Robert A. Milne, Hassan Barada, Mohammed Al-Muallia, Arif Al Hammadi, and Tod Laursen for their thoughtful review of the original project proposal and for their support during the pilot implementation. We also thank Marianne Hassan for her support, particularly for her efforts to organize institutional ethics review of the project. This research is internally supported by the Office of Research Support, Khalifa University.

*george.hitt@kustar.ac.ae

1. L. C. McDermott and E. F. Redish, Am. J. Phys. 67, 755 (1999).
2. R. R. Hale, Am. J. Phys. 66, 64 (1998).
3. J. W. Belcher, The MIT Faculty Newsletter 16, 8 (2003).
4. N. D. Finkelstein and S. J. Pollock, Phys. Rev. ST Phys. Educ. Res. 1, 010101 (2005).
5. N. D. Finkelstein, Int. J. Sci. Educ. 27, 1187 (2005).
6. C. Henderson and M. H. Dancy, Phys. Rev. ST Phys. Educ. Res. 3, 020102 (2007).
7. M. Dancy and C. Henderson, Am. J. Phys. 78, 1056 (2010).
8. P. Heller, R. Keith, and S. Anderson, Am. J. Phys. 60, 627 (1992).
9. P. Heller and M. Hollabaugh, Am. J. Phys. 60, 637 (1992).
10. P. Heller and K. Heller, Cooperative Group Problem Solving in Physics (University of Minnesota, http://groups.physics.umn.edu/physed/Research/CGPS/, 1999) last accessed 10 March 2012.
11. F. Goldberg, V. Otero, and S. Robinson, Am. J. Phys. 78, 1265 (2010).
12. E. F. Reddish, J. M. Saul, and R. N. Steinburg, Am. J. Phys. 66, 212 (1998).
13. D. Hestenes, M. Wells, and G. Swackhammer, Phys. Teach. 30, 141 (1992).
14. International English Language Testing System (IELTS) test. http://www.ielts.org/about_us.aspx. Last accessed 6 Dec. 2012.
15. The Road not Traveled: Education Reform in the Middle East and Africa (The World Bank, Washington DC, USA, 2008).
16. F. Heard-Bey, From Trucial States to United Arab Emirates, illustrated ed. (Motivate Publishing Ltd., Dubai, UAE, 2005).
17. C. Davidson, in Higher Education in the Gulf States: Shaping Economies, Politics, and Culture, edited by C. Davidson and P. M. Smith (London Middle East Institute at SAOS, London, UK, 2008).
18. N. Ridge, “The middle east student in higher education,” private communications (2012), Dr. N. Ridge is Executive Director of the Sheikh Saud bin Saqr Al Qasimi Foundation for Public Policy Research, a public policy think tank commissioned by the northern Emirate of Ras al Khaimah, which focuses on research-driven education reform studies of Ras al Khaimah and the UAE educational systems.
19. D. Demaree, S. Allie, M. Low, and J. Taylor, in Physics Education Research Conference 2008, PER Conference, Vol. 1064 (Edmonton, Canada, 2008) pp. 107–110.
20. The central argument to attend college, in advocacy to US high school students, is invariably financial and frequently cites studies like those from the Advocacy & Policy Center of CollegeBoard (which maintains the SAT). Its highly cited ”EducationPays" series tracks 3-year averages in the earnings of college graduates. See http://trends.collegeboard.org/educationpays. Last accessed 3 May 2012.
21. J. D. H. Gaffney, A. L. H. Gaffney, and R. J. Beichner, Phys. Rev. ST Phys. Educ. Res. 6, 010102 (2010).
22. J. Baron, in ERGA Conference (2008).
23. K. Litvin and M. McAllister, “Understanding the cultural dimensions of Gulf Arab students,” (2011), unpublished.
24. C. Turpen and N. D. Finkelstein, Phys. Rev. ST Phys. Educ. Res. 6, 020123 (2010).
25. G. Hofstede, Culture’s Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations, 2nd ed. (Sage Publications, Inc., Los Angeles, USA, 2001).
26. G. Hofstede, (2001), "Power distance" is the construct described by G. Hofstede as "the extent to which the less powerful members of organizations, institutions, and families accept and expect authority to be distributed unequally". On a 100-point scale, American culture is scored at 40 as of 2011 which compared to other nations is relatively low. Gulf Arab culture (excluding Saudi Arabia) scores 90 which is relatively very high. The expectations listed in the text are correlated with these scores and are taken from Figure 3.8, pp. 107 of G. Hofstede (2001).
27. G. Hofstede, (2001), "Uncertainty avoidance" is the construct described by G. Hofstede as "society’s tolerance for ambiguity". On a 100-point scale, American culture
is scored at 46 as of 2011 which is relatively low. Gulf Arab culture (excluding Saudi Arabia) scores 80 which is relatively high. The Arab culture (excluding Saudi Arabia) scores 25 which is relatively low. The expectations listed in the text are correlated with these scores and are taken from Figure 4.6, pp. 169 of G. Hofstede (2001).

28 G. Hofstede, (2001), "Individualism" is the construct described by G. Hofstede as "the degree to which individuals are integrated into groups". On a 100-point scale, American culture is scored at 91 as of 2011 which is relatively very high. Gulf Arab culture (excluding Saudi Arabia) scores 25 which is relatively very low. The expectations listed in the text are correlated with these scores and are taken from Figure 5.5, pp. 236 of G. Hofstede (2001).

29 I. Halloun and D. Hestenes, Sci. Educ. 7, 533 (2005).

30 W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, Phys. Rev. ST Phys. Educ. Res. 2, 010101 (2006).

31 H. W. Stevenson and J. W. Stigler, The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education (Simon & Schuster, New York, USA, 1994).

32 M. A. Runco, Divergent Thinking (Ablex Publishing, New York, USA, 1991).

33 G. Land and B. Jarmon, Breakpoint and Beyond (Leadership 2000 Inc, Scottsdale, USA, 1998).

34 Values Survey Module version 1994 (VSM94), http://www.geerthofstede.nl/vsm-94. Last accessed 3 Jan. 2013.

35 T. A. Laursen et al., Building a community of excellence: Khalifa University strategic plan 2012-2017, Tech. Rep. (Khalifa University, 2012) town hall discussions were major inputs for a recent strategic planning exercise. Available at http://www.kustar.ac.ae/aboutus/pdf/Strategic_Document_2012-2017.pdf.

36 D. S. Domin, J. Chem. Educ. 76, 543 (1999).

37 L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Am. J. Phys. 72, 1147 (2004).

38 M. Hollander and D. A. Wolfe, Nonparametric Statistical Methods, 2nd ed. (Wiley-Interscience, New York, USA, 1999).

39 ETS, ETS Research Reports (2010).

40 T. R. Khan, “A pre-calculus basic skills test,” (2011), unpublished.

41 J. R. Taylor, An Introduction to Error Analysis, 2nd ed. (University Science Books, Sausalito, CA, 1997).

42 C. J. Atman et al., Enabling Engineering Student Success: The Final Report for the Center for the Advancement of Engineering Education (Morgan & Claypool, San Rafael, CA, 2010).

43 R. R. Hake, Am. J. Phys. 66, 64 (1998), this value for Hake’s gain for U.S. engineering students is calculated using only his data for calculus-based college and university physics courses. In his original paper, there are 7 such courses (gains); M92-C (0.22), ASU1-C (0.25), ASU2-C (0.23), CP-C (0.25), EM90-C (0.27), OS92-C (0.13), and UL94F-C (0.18). The population size for these is N = 1194.

44 R. J. Beichner, “The scale-up project: A student-centered active learning environment for undergraduate programs,” (2008), paper presented at the National Research Council Workshop on Linking Evidence to Promising Practices in STEM Undergraduate Education, Washington, DC. Available: http://www7.nationalacademies.org/bose/Beichner_CommissionedPaper.pdf.

45 R. A. Goodman, Phys. Teach. 16, 367 (1978).

46 R. R. Hake, “Interactive-engagement methods in introductory mechanics courses,” (1998), includes 6 cases studies of various secondary implementations of IE teaching techniques that report low normalized gain from pre-/post- FCI. Instructors in each case report various implementation problems. Available at http://www.physics.indiana.edu/~sdi/IEM-2b.pdf. Last accessed 9 April 2012. (unpublished).

47 C. Turpen and N. Finkelstein, Phys. Rev. ST Phys. Educ. Res. 5, 020101 (2009).

48 One author, G. W. Hitt, was a graduate teaching assistant at Michigan State Univ. in academic year 2003/04, where the introductory physics courses use the LON-CAPA online homework and discussion system.

49 One author, A. F. Isakovic, was a graduate teaching assistant at Univ. of Minnesota in academic year 1998/99, where he also attended the formal training in CGPS. He later informally implemented CGPS in calculus-based general physics while postdoctoral research associate and lecturer at Cornell University in Spring and Fall 2006.

50 G. M. Novak, E. T. Patterson, A. D. Gavrin, and W. Christian, Just-in-time Teaching: Blending Active Learning with Web Technology (Prentice Hall, Englewood Cliffs, NJ, 1999).

51 D. Hestenes, Am. J. Phys. 55, 440 (1987).

52 E. Mazur, Peer Instruction, A User’s Manual (Prentice Hall, Upper Saddle River, NJ, USA, 1997).

53 R. J. Beichner, in Research-Based Reform of University Physics, Vol. 1 (http://www.compadre.org/Repository/document/ServeFile.cfm?ID=4517&DocID=183, 2007) last accessed 9 April 2012.

54 R. R. Hake, Phys. Teach. 30, 546 (1992).

55 L. C. McDermott and P. S. Shaffer, Tutorials in Introductory Physics, 1st ed. (Prentice Hall, Upper Saddle River, NJ, USA, 2002).

56 M. C. Wittmann, R. N. Steinberg, and E. F. Redish, Activity-Based Tutorials: Introductory Physics, The Physics Suite, 1st ed., Vol. 1 (Wiley, Upper Saddle River, NJ, USA, 2004).

57 R. E. Scherr and A. Elby, in AIP Conf. Proc., Vol. 883, edited by P. R. L. Heron, L. McCullough, and J. Marx (2006) pp. 46-49.

58 P. W. Laws, Physics Today 44, 24 (1991).

59 P. W. Laws, Activity-Based Tutorials: Introductory Physics, The Physics Suite, 1st ed., Vol. 1 (Wiley, Upper Saddle River, NJ, USA, 2004).

60 R. R. Hake, See http://www.physics.indiana.edu/~sdi/. Last accessed 9 April 2012. The SDI lab curriculum is significantly narrower than the traditional introductory lab curriculum and focuses exclusively on experiments with kinematics and Newton’s Laws.

61 E. Etkina, K. Gibbons, B. L. Holton, and G. K. Horton, Am. J. Phys. 67, 810 (1999).

62 E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O’Brien, I. Rodriguez, and P. Pemala, Phys. Rev. ST Phys. Educ. Res. 6, 010106 (2010).

63 G. Kortemeyer, Phys. Rev. ST Phys. Educ. Res. 3, 010101 (2007).

64 K. Malone, Phys. Rev. ST Phys. Educ. Res. 3, 020101 (2008).

65 N. Lasry, E. Mazur, and J. Watkins, Am. J. Phys. 76, 1066 (2008).

66 P. Heller and K. Heller, Physics 1301 Laboratory Manual (University of Minnesota,
http://groups.physics.umn.edu/physted/Research/Lablast accessed 7 April 2012., 2007) last accessed 7 April 2012.

67 K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Phys. Educ. Res., Am. J. Phys. Suppl. 67, S38 (1999).

68 S. P. Formica, J. L. Easley, and M. C. Spraker, Phys. Rev. ST Phys. Educ. Res. 6, 020106 (2010).

69 A. P. Fagen, C. H. Crouch, and E. Mazur, Phys. Teach. 40, 206 (2002).

70 L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Phys. Rev. ST Phys. Educ. Res. 5, 010101 (2009).

71 J. M. Saul and E. F. Redish, “Final evaluation report for FIPSE grant #P116P50026: Evaluation of the workshop physics dissemination project.” (1997), available at http://www.physics.ucf.edu/~saul/articles/WP-FIPSE_Rprt.pdf. Last accessed 1 March 2011.

72 E. Brewe, L. H. Kramer, and G. E. O’Brien, Phys. Rev. ST Phys. Educ. Res. 5, 013102 (2009).

73 A "low-effort" designation is given to those IE pedagogies that were designed by starting with all the features of traditional instruction, iteratively making small changes, and converging on an effective model that still has many structural similarities with traditional instruction. In contrast, "high-effort" designation is given to those IE pedagogies that were design by a "clean-slate" approach, where all course elements were created from scratch, have course elements with no analog in the traditional format, require specialized technical support staff, or where substantial training is required before attempting to teach the new method.

74 See http://modeling.asu.edu, Modeling Instruction Workshops. Modeling instruction is judged to be a "high effort" pedagogy because of the intensive, specialized training needed, even to implement MI at the high school level.

75 J. P. Kotter, Harvard Business Review March-April (1995).

76 C. L. Dym, A. A. Agogino, O. Eris, D. D. Frey, and L. J. Liefer, J. Eng. Educ. 94, 103 (2005).

77 M. T. H. Chi, P. J. Feltovich, and R. Glaser, Cog. Sci. 5, 121 (1981).

78 J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. (Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1988).

79 M. Lorenzo, C. H. Crouch, and E. Mazur, Am. J. Phys. 74, 118 (2006).

80 K. Kreutzer and A. Boudreaux, Phys. Rev. ST Phys. Educ. Res. 8, 010120 (2012).

81 R. M. Goertzen, E. Brewe, L. H. Kramer, L. Wells, and D. Jones, Phys. Rev. ST Phys. Educ. Res. 7, 020105 (2011).