New Orleans Jazzed about Engineering Education

Paper ID #15591

Analyzing an Abbreviated Dynamics Concept Inventory and Its Role as an Instrument for Assessing Emergent Learning Pedagogies

Mr. Nick Stites, Purdue University, West Lafayette

Nick Stites is pursuing a PhD in Engineering Education at Purdue University. His research interests include the development of novel pedagogical methods to teach core engineering courses and leveraging technology to enhance learning experiences. Nick holds a BS and MS in Mechanical Engineering and has eight years of engineering experience. He also has four years of experience as an adjunct instructor at the community-college and research-university level.

David A Evenhouse, Purdue University

David Evenhouse is a Graduate Student and Research Assistant in the Purdue School of Engineering Education. He graduated from Calvin College in the Spring of 2015 with a B.S.E. concentrating in Mechanical Engineering. Experiences during his undergraduate years included a semester in Spain, taking classes at the Universidad de Oviedo and the Escuela Politécnica de Ingenieria de Gijón, as well as multiple internships in Manufacturing and Quality Engineering. His current work primarily investigates the effects of select emergent pedagogies upon student and instructor performance and experience at the collegiate level. Other interests include engineering ethics, engineering philosophy, and the intersecting concerns of engineering industry and higher academia.

Mariana Tafur, Purdue University, West Lafayette

Mariana Tafur is an assistant professor at University of Los Andes in Bogotá - Colombia. She has a Ph.D. in Engineering Education at Purdue University, West Lafayette-IN; a M.S., in Education at Los Andes University, Bogotá-Colombia; and a B.S., in Electronics Engineering at Los Andes University, Bogotá-Colombia. She is a 2010 Fulbright Fellow. Her research interests include engineering skills development, STEM for non-engineers adults, motivation in STEM to close the technology literacy gap, STEM formative assessment, and Mixed-Methods design.

Prof. Charles Morton Krousgrill, Purdue University, West Lafayette

Charles M. Krousgrill is a Professor in the School of Mechanical Engineering at Purdue University and is affiliated with the Ray W. Herrick Laboratories at the same institution. He received his B.S.M.E. from Purdue University and received his M.S. and Ph.D. degrees in Applied Mechanics from Caltech. Dr. Krousgrill's current research interests include the vibration, nonlinear dynamics, friction-induced oscillations, gear rattle vibrations, dynamics of clutch and brake systems and damage detection in rotor systems. Dr. Krousgrill is a member of the American Society for Engineering Education (ASEE). He has received the H.L. Solberg Teaching Award (Purdue ME) seven times, A.A. Potter Teaching Award (Purdue Engineering) three times, the Charles B. Murphy Teaching Award (Purdue University), Purdue's Help Students Learn Award, the Special Boilermaker Award (given here for contributions to undergraduate education) and is the 2011 recipient of the ASEE Mechanics Division's Archie Higdon Distinguished Educator Award.

Craig Zywicki, Purdue University

Craig is a Data and Assessment Analyst in the Office of Institutional Research, Assessment, and Effectiveness at Purdue University.

Dr. Angelika N Zissimopoulos, University of Chicago

Angelika Zissimopoulos holds a Ph.D. in Biomedical Engineering From Northwestern University. She is currently the Associate Director for STEM education at the University of Chicago.

Dr. David B Nelson, Purdue University, West Lafayette

David B. Nelson is Associate Director of the Center for Instructional Excellence at Purdue University. He received his Ph.D in World History from the University of California, Irvine in 2008.

David has been involved in many educational research projects at Purdue, including published worked in the programming education, student engagement and academic performance in dynamics engineering courses, and educational modalities in engineering, technology and economics.

Prof. Jennifer DeBoer, Purdue University, West Lafayette

Jennifer DeBoer is currently Assistant Professor of Engineering Education at Purdue University. Her research focuses on international education systems, individual and social development, technology use and STEM learning, and educational environments for diverse learners.

Prof. Jeffrey F Rhoads, Purdue University, West Lafayette

Jeffrey F. (Jeff) Rhoads is an Associate Professor in the School of Mechanical Engineering at Purdue University and is affiliated with both the Birck Nanotechnology Center and Ray W. Herrick Laboratories at the same institution. He received his B.S., M.S., and Ph.D. degrees, each in mechanical engineering, from Michigan State University in 2002, 2004, and 2007, respectively. Dr. Rhoads' current research interests include the predictive design, analysis, and implementation of resonant micro/nanoelectromechanical systems (MEMS/NEMS) for use in chemical and biological sensing, electromechanical signal processing, and computing; the dynamics of parametrically-excited systems and coupled oscillators; the behavior of electromechanical and thermomechanical systems, including energetic materials, operating in rich, multi-physics environments; and mechanics education. Dr. Rhoads is a member of the American Society for Engineering Education (ASEE) and the American Society of Mechanical Engineers (ASME), where he serves on the Design, Materials and Manufacturing Segment Leadership Team and the Design Engineering Division's Technical Committees on Micro/Nanosystems and Vibration and Sound. Dr. Rhoads is a recipient of the National Science Foundation's Faculty Early Career Development (CAREER) Award, the Purdue University School of Mechanical Engineering's Harry L. Solberg Best Teacher Award (twice), and the ASEE Mechanics Division's Ferdinand P. Beer and E. Russell Johnston, Jr. Outstanding New Mechanics Educator Award. In 2014, Dr. Rhoads was selected as the inaugural recipient of the ASME C. D. Mote Jr., Early Career Award and was featured in ASEE Prism Magazine's 20 Under 40.

Dr. Edward J. Berger, Purdue University, West Lafayette

Edward Berger is an Associate Professor of Engineering Education and Mechanical Engineering at Purdue University, joining Purdue in August 2014. He has been teaching mechanics for nearly 20 years, and has worked extensively on the integration and assessment of specific technology interventions in mechanics classes. He was one of the co-leaders in 2013-2014 of the ASEE Virtual Community of Practice (VCP) for mechanics educators across the country.

Analyzing an Abbreviated Dynamics Concept Inventory and Its Role as an Instrument for Assessing Emergent Learning Pedagogies

Abstract

The Dynamics Concept Inventory (DCI) is a validated assessment tool commonly used to evaluate student growth within core, gateway-level mechanics courses. This research explored the evaluative use of this tool within the context of Freeform – an emergent course system that buttresses active class meetings with blended and collaborative virtual learning environments, themselves founded upon extensive multimedia content and interactive forums – at Purdue University. The paper specifically considers a number of related issues including: (i) the thoughtful development (via expert content validation) and statistical reliability of an abbreviated DCI instrument, which is more amenable to in-class implementation than the much longer full DCI; (ii) the correlation of abbreviated-DCI performance with exam scores and final course grades for a dynamics course using the Freeform framework with an emphasis on both conceptual understanding and traditional problem-solving skills; and (iii) various inter-section performance metrics in a preliminary study on how an implementation of the abbreviated-DCI may help elucidate the impact of the instructor within the Freeform framework. The results of these analyses supported the validity and reliability of the abbreviated DCI tool, and demonstrated its usefulness in a formal research setting. The preliminary study suggested that the Freeform framework might normalize differences in instructor pedagogical choices and student performance across class sections. These findings indicate that the abbreviated DCI holds promise as a research instrument and lay the groundwork for future inquiry into the impact of the Freeform instructional framework on students and instructors alike.

Introduction

Undergraduate students often find dynamics to be a challenging, gateway engineering course[1]. Typically offered at the sophomore level, dynamics combines many fundamental concepts from physics, calculus, and statics to build a foundation for many higher-level engineering courses. Unfortunately, dynamics has been plagued historically by a large number of students earning a D, an F, or withdrawing from the course (the course's "DFW" rate). A DFW rate of over 20% for a dynamics class at Purdue University prompted two professors to develop and implement a new instructional framework in 2008[2]. This new framework, labeled *Freeform*, incorporates active, blended, and collaborative learning environments both in class and online. Since the inception and implementation of Freeform, the DFW rate for dynamics at Purdue has dropped to near 10%, which translates to approximately 50 more students passing dynamics per year.

The Freeform philosophy incorporates a balanced emphasis of both traditional problem solving skills and conceptual understanding, all supported by a robust range of online learning resources. The Freeform *lecturebook* is a hybrid of a traditional textbook and lecture notes tailored to the Freeform environment and is inspired by the workbooks commonly used in primary education. It includes short-answer and example problems (to practice problem-solving techniques), multiple-choice and short-answer conceptual problems, and ample space for in-class note taking. Every

example problem in the lecturebook is paired with an online instructional video demonstrating the problem solution. The online video library also includes video solutions for every homework problem and conceptual-demonstration videos. These online resources are available to all students through the course blog, which also facilitates asynchronous communication via threaded discussions of homework problems or exam content. At Purdue, the lecturebook, homework assignments, exams, and availability of online content are consistent across all sections of dynamics, but each instructor is empowered to employ whatever in-class pedagogies, examples, and quizzes they deem appropriate. Thus, Freeform presents instructors with a de-centralized (in terms of pedagogical decision making) but highly-supported (in terms of the lecturebook, online community, and content) environment in which to teach. In essence, Freeform is a strongly-scaffolded learning system that empowers individual instructors with pedagogical choice and individual students with a variety of learning opportunities and support.

Freeform's attempt to balance conceptual understanding with problem solving ability also directly promotes student competency in both areas. Each instructor encourages students to develop a conceptual understanding of main class topics through the use of in-class activities, discussion of conceptual questions, or the utilization of available online resources. Therefore, at Purdue, conceptual problems often constitute a significant portion of the final exam grade for all sections, as well as a significant portion of credit on the three midterm exams.

The Freeform environment represents a significant shift from traditional pedagogical practice in dynamics. Therefore, we have embarked upon an extensive evaluation of Freeform to isolate and assess its effects on student learning. One area of interest involves the evaluation of the students' conceptual understanding of dynamics, as this is one of Freeform's core areas of emphasis. To this end, a pre-existing assessment tool geared to measure student conceptual understanding in the field of dynamics, referred to as the Dynamics Concept Inventory (DCI)[3], was considered. However, to streamline the implementation of such an instrument, the Purdue instructors wished to incorporate the concept inventory into the final exam as a replacement for the conceptual questions already asked as part of this evaluation. Concerned that the entire 29-question DCI was too long for the given environment, subject-matter experts selected an 11-question subset of the DCI that still provided a comprehensive assessment of conceptual understanding given the constraints of the final exam format. While serving as a portion of the final exam, this abbreviated DCI could also function as an instrument for both cross-sectional and longitudinal research studies of student conceptual performance within the Freeform environment.

This paper discusses the development of the aforementioned abbreviated DCI (aDCI), including the evaluation of its validity and reliability, and the role that it can play within broader, complex studies of the Freeform instructional framework. This work also includes a preliminary study in which the results of the aDCI are evaluated alongside a number of traditional assessment metrics in order to begin exploring instructor effects within the Freeform framework. A preliminary hypothesis is that the Freeform framework reduces variations in student performance due to differences in instructional choices or style through its extensive scaffolding and support resources, thus providing a more consistent student experience across multiple sections of the same course. While it is not expected that this one preliminary study will fully answer the research question posed above, the analysis and results presented here will inform and guide future inquiry.

Background

Over the past two decades there has been a distinct rise in the popularity of concept inventories (CIs) for use in education research. This is directly attributable to the need for valid and reliable methods to assess student conceptual understanding and for instruments that can capture data on both the misunderstandings students hold and their general comprehension of key concepts[4]. The Dynamics Concept Inventory (DCI) was released for public use in 2005 and was the result of a multi-year process involving a Delphi study, student focus groups, and extensive beta testing[3,5]. The test itself is designed to address 11 distinct concepts and is comprised of 29 questions, four of which were taken from the pre-existing Force Concept Inventory[3]. It was developed, like many CIs, with the intention of providing a valid and reliable instrument, capable of evaluating the effects of innovative or experimental instructional practices upon student learning[3,4].

Instructors at Purdue decided to supplant the existing conceptual questions on their final exam with DCI questions. An abbreviated version of the DCI had to be developed and utilized because the 29-question version was considered too long for the time dedicated to conceptual questions on the final exam. The incorporation of an abbreviated DCI into the final negated the need to sacrifice another full class period for administering the CI post-test. It should be noted that neither the process of integrating a CI into an assessment for a course, nor the abbreviation of a research tool, is without precedent. For example, Smith, Wood, and Knight incorporated a genetics concept inventory into the final exam[6]. Additionally, Henderson studied if grading a concept inventory significantly altered student performance and found that students will put forth an honest effort on concept inventories regardless of the incentives involved[7]. Henderson's result helps justify the Purdue instructors' decision to provide students with a completion grade for the pre-test (researchers scored the tests independently of the class) and to incorporate the post-test into the final examination.

The reduction of the number of CI questions in an effort to lower the required completion time has also been explored previously. A general example would be the Big 5 Factors personality inventory for which many implementations have dozens of items. A shorter, 10-item version of the Big 5 inventory has been validated for use with a variety of applications specifically because users of the inventory (as opposed to personality researchers or inventory developers) need a tool that balances its length with accuracy[8,9]. Likewise, Han et al. conducted split-half reliability tests to determine the internal reliability of the Force Concept Inventory[10]. The authors divided the 30-question FCI test into two equal 15-question halves and administered each half and the unaltered full test as its own instrument. In the analysis of any two instruments, the mean error of equivalent scores between the two differed by only 3%[10], indicating that the halved CI could be used as a reliable instrument for score-based assessment.

While it is common practice for CI development teams to publish articles concerning the design and validation of their tests[3,11], a number of researchers have begun to independently examine CIs in light of Classical Test Theory (CTT), Item Response Theory (IRT), or other evaluative frameworks[12–14]. This psychometric analysis allows for independent verification of a CI's performance in the classroom and its utility as a research tool. A few of these studies have

specifically evaluated the DCI, notably the works of Jorion et al.[12,14,15]. Their studies have included evaluation of internal reliability using Cronbach's alpha coefficient[12], examining item reliability using measures of difficulty and discrimination[15], and factorial evaluation based on an expanded 16-concept categorical scheme that the DCI authors created by modifying their original 11 concept categories[14]. Based on their findings, Jorion et al. suggested that the application of the DCI be limited to low-stakes research environments analyzing class-aggregate data[12]. According to guidelines from CTT, the DCI is not suitably reliable to allow for statistical analysis on an individual student basis[13]. Additionally, factorial analysis of student responses did not support the original 11-concept or the 16-concept structure defined by the DCI creators. This implies that the test would require further refinement in order to reveal meaningful data concerning student comprehension of specific concepts. However, it is still reliable for looking at student conceptual understanding of dynamics as a whole. These recommendations, when combined, provide an indication of the role that the DCI or the abbreviated DCI (aDCI) should take when applied to the evaluation of both student outcomes and the Freeform framework. Because the aDCI is a new tool made from the sufficiently valid and reliable DCI, only basic psychometric evaluation of the aDCI will be included in this paper.

Explanation of Data Collection and Preliminary Study

The data for the psychometric testing of the aDCI were collected at Purdue during the Spring 2015 semester. The dataset included 361 complete responses (i.e., complete pre- and post-tests) from students distributed among three sections of dynamics taught within the university's Mechanical Engineering department. These three sections also provide the data for the preliminary study which utilizes the aDCI to assess the influence of instructional differences on student performance within the Freeform framework. Most of the students included in the study were in their second year of undergraduate coursework, and the vast majority (72%, 94%, and 88% for sections 1, 2, and 3, respectively) were Mechanical Engineering students. The three instructors of the different sections all had prior experience teaching dynamics within the Freeform framework. Each of the sections had common homework assignments, midterm exams, final exams, and course policies defined in the course syllabus. The three sections also shared a common blog space for online collaboration and communication. However, each instructor had the freedom to use their own pedagogical discretion in planning class activities and assigning quizzes. During the second week of classes, the pre-test of the 11-item aDCI was administered in a pencil-and-paper format during class. The identical aDCI post-test was incorporated into the final exam. Because the 11 questions of the aDCI were chosen in part because they aligned with the content taught at Purdue, the aDCI was an appropriate summative assessment of conceptual knowledge. Also, it should be noted that the conceptual questions in the lecturebook probe the same concepts covered on the aDCI (and more), but the questions in the lecturebook are *not* duplicated from the aDCI—the lecturebook questions were published before implementing the aDCI as part of this course. Finally, as mentioned previously, the inclusion of conceptual problems on the final exam was standard practice for the dynamics instructors at Purdue. Thus, the aDCI simply filled this pre-existing role.

Development of the aDCI and Psychometric Evaluation

Question Selection and Expert Validation

Previous publications regarding the development of the DCI included a taxonomy of 11 underlying concepts covered by the 29 conceptual questions[3]. However, the authors did not explicitly state which questions they considered to belong in each conceptual category. Therefore, two subject-matter experts in, and veteran instructors of, dynamics at Purdue combined efforts to sort the 29 DCI questions into the 11 pre-defined conceptual categories. These experts then carefully selected 11 questions to comprise the aDCI that would span all 11 conceptual categories of the DCI and would align well with the established dynamics curriculum at Purdue.

The selection of more than one question from each concept category would have been preferable in order to more reliably test a student's understanding of the concept; however, 22 questions (two questions per concept) was deemed unreasonable to include as part of the final exam. Thus, only 11 questions were selected. The 11 questions included in the aDCI are cross-referenced to their original DCI question number in Table 1.

Table 1. The 11-question subset that formed the aDCI.

Abbreviated DCI Question #	1	2	3	4	5	6	7	8	9	10	11
Original DCI Question #	1	4	7	18	9	10	11	13	19	21	22

Exploring Validity by Comparing aDCI Scores, Exam Scores, and Final Grades

The concurrent and convergent validity of the aDCI were evaluated via comparison of the aDCI scores, final exam scores, and final grades. This comparison is a very practical and easy-to-implement approach. However, there is part-whole dependency between aDCI scores and final exam grades because the aDCI was embedded into the final exam. To address this dependency, the aDCI questions were removed from the scoring of the final exam in a process similar to the one employed by Smith et al.[6] Final exam and final grade percentages were recalculated to reflect the exclusion of the aDCI in a new scoring method. As a consequence of removing the conceptual questions, the final exam is weighted almost exclusively toward the assessment of problem-solving skills, rather than conceptual understanding. Additionally, final exam scores constitute 25% or 50% (depending on final-exam performance relative to the midterm-exam average) of a student's final grade, producing another significant relationship between variables. However, previous research suggests that conceptual knowledge and problem-solving skills could correlate significantly to one another[16,17]. Therefore, although not ideal, the correlation between the aDCI, exam scores, and final grade was considered sufficient for the scope and purpose of this analysis.

The correlations between the aDCI scores, exam scores, and final grade in the class (by percentages) are presented in Table 2. The normalized gain metric is defined in the same manner as the metric utilized by Hake when examining FCI performance[18], but is applied to individual scores rather than class averages. The governing equation is:

$$G = \frac{Post\ Score\ \% - Pre\ Score\ \%}{100 - Pre\ Score\ \%}, \qquad (1)$$

where G is the normalized gain[19]. The gain metric can be thought of as how much conceptual understanding the student gained relative to the maximum-possible gain for that student.

The aDCI scores are ordinal data, so the Spearman rho correlation coefficient was utilized to quantify the associations between variables. While the correlations are not as strong as those reported by Smith and colleagues[6], they are statistically significant and similar to those reported by Steif for the Statics Concept Inventory[20]. Additionally, the moderate correlation coefficients between the inventory scores and exam scores fall in the range of values found in previous publications comparing concept scores to problem-solving skills[16]. This fits with the observation that much of the final grade and the exam scores reflect assessments of problem-solving rather than conceptual understanding. Overall, the expert selection of questions for the 11-question subset and the significant correlations between the aDCI scores and other assessment metrics provide evidence that the aDCI is sufficiently valid for use in this study.

Table 2. Spearman correlation coefficient (ρ) for aDCI scores and other performance metrics.

	aDCI Pre-Test Scores		aDCI Post-Test Scores		aDCI Gain, *G*	
	ρ	SE	ρ	SE	ρ	SE
aDCI Post-Test Scores	0.57***	0.043				
aDCI Gain, *G*	0.14**	0.052	0.87***	0.026		
Exam 1 Scores	0.18***	0.052	0.37***	0.049	0.35***	0.050
Exam 2 Scores	0.37***	0.049	0.44***	0.047	0.34***	0.050
Exam 3 Scores	0.29***	0.051	0.43***	0.048	0.37***	0.049
Final Exam Score (no aDCI)	0.13*	0.052	0.29***	0.051	0.30***	0.050
Final Grade % (no aDCI)	0.22***	0.052	0.44***	0.048	0.43***	0.048

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

aDCI Reliability

Two principle methods were employed to evaluate the reliability of the aDCI as an instrument to assess conceptual understanding. Cronbach's alpha measured the internal reliability of the test as a whole, and the alpha-when-item-deleted metric determined if any questions abnormally lowered the value of Cronbach's alpha.

The overall Cronbach alpha for the aDCI was 0.475 and 0.607 for the pre- and post-tests, respectively. The increase in alpha from the pre- to post-test is indicative of the class as a whole answering the questions more consistently and a likely artifact of reduced instances of guessing on the post-test[21]. The post-test alpha of 0.607 is lower than those published by Gray et al. (0.640-0.837) and the team led by Jorion (0.74) for the full 29-question DCI[3,12]. However, Cronbach's alpha is directly dependent on the number of questions on a given test, so it is difficult to determine if the difference in alpha between the aDCI and the DCI is mostly due to the difference in exam length or other causes[21].

Regardless of the cause of the moderately low alpha on the post-test, an alpha of 0.607 is less than what some researchers recommend for high-stakes testing[13,15]. This is somewhat concerning, as performance on the aDCI could account for up to 25% of a student's overall course grade. Nevertheless, it is important to remember that conceptual questions were part of the final exam prior to developing the aDCI, and it is very probable, but not easily proven, that the conceptual question sets made for the final each year have an even lower internal reliability than that demonstrated by the aDCI. Thus, it is highly likely that the use of the aDCI on the final exam is actually benefiting the students, rather than jeopardizing their performance.

Because alpha is dependent on the number of questions, it is expected that when a question is removed from the alpha calculation that the new alpha, the *alpha-with-item-deleted*, will decrease. If the alpha-with-item-deleted increases, the question that was deleted is abnormally decreasing the alpha of the overall test, and the item should be considered for modification or replacement[21]. Only one item, Question 6, shown in Figure 1, was found to have an alpha-with-item-deleted (0.635) higher than the overall alpha. Jorion et al. also identified this question as having a higher alpha-with-item-deleted, thus, Question 6 is a strong candidate for modification or replacement in both the full DCI and the aDCI[12].

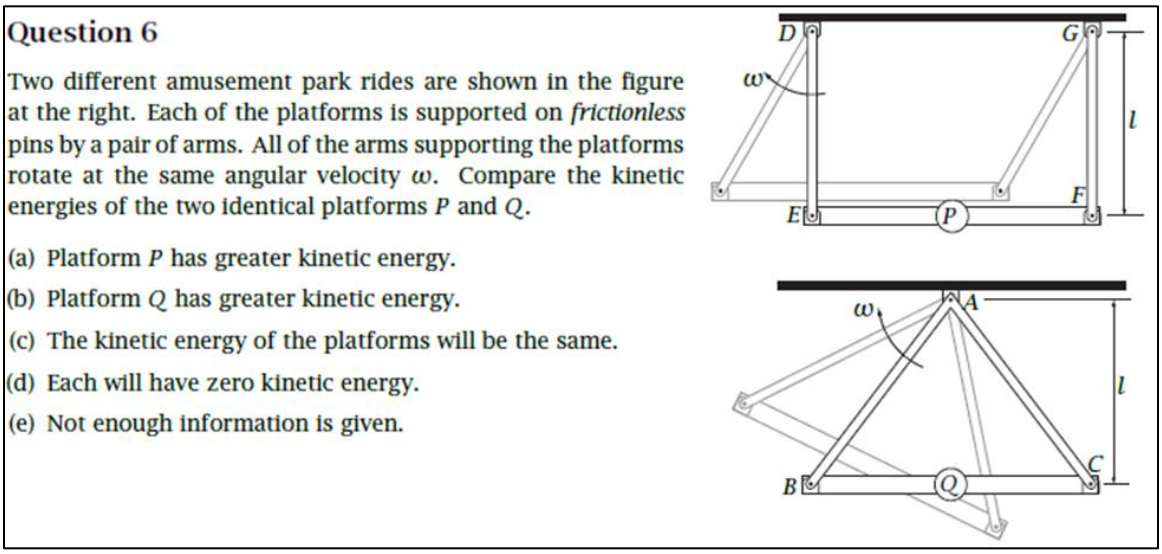


Figure 1. Question 6 of the aDCI had a higher alpha-with-item-deleted than the Cronbach's alpha of the entire test, indicating it is a candidate for modification or replacement (adapted from Gray et al.[3]).

aDCI Item Performance

The item difficulty for each question of the aDCI was used to identify too-easy or too-hard questions, and the item discrimination of each question was used to assess how well the item differentiated high-performing and low-performing students. Similar to the alpha-when-item-deleted metric, the main objective of calculating these measures was to identify specific questions for modification or replacement.

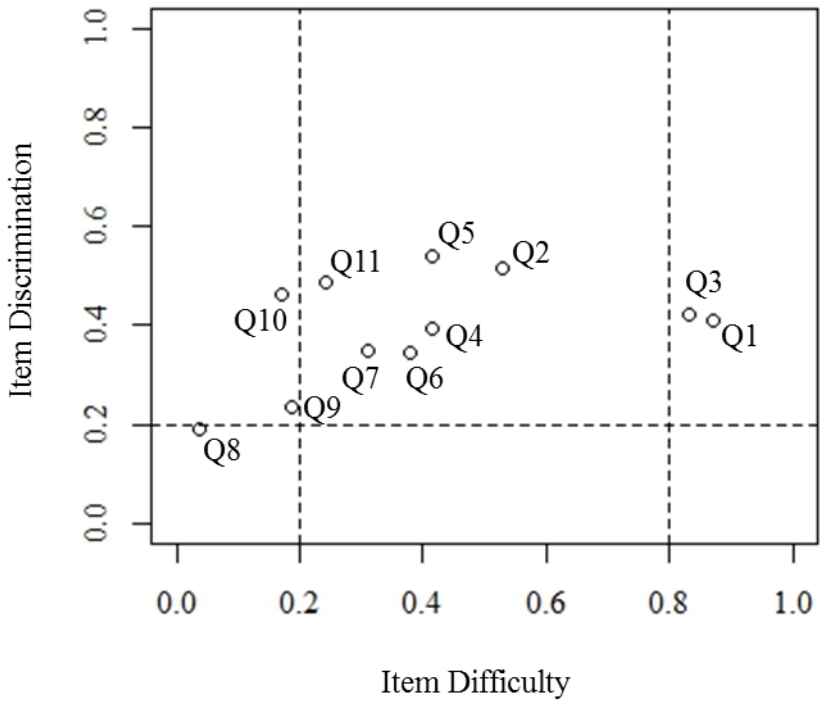
The item difficulty for each of the 11 questions was calculated with the definition of difficulty previously used by other DCI researchers—the item difficulty is simply the proportion of students answering a given question correctly[13,15]. Commonly, an item difficulty between 0.2 and 0.8 indicates that the question is neither too easy nor too hard[13,15]. The item difficulties for the aDCI and the suggested thresholds (dashed lines) are displayed in Figure 2. A general shift upward in item difficulty between the pre- and post-tests indicates that more students answered questions correctly, implying a gain in conceptual understanding. Greater than 80% of the students answered Questions 1 and 3 correctly on the pre- and post-tests, but these questions are known to be review material, as they were taken from the FCI by the original creators of the DCI[5]. They are specifically included to validate the assumption that students have prior knowledge in Newtonian mechanics. Question 11 joined Questions 1 and 3 above the 80% difficulty threshold on the post-test. This question pertains to the speeds of different locations of a rolling-without-slipping wheel, and Grey et al. reported similarly high correct-response rates (near 77%) on post-tests[3]. Overall, no aDCI items performed unusually with regard to item difficulty.

Item discrimination details how well an item differentiated between high-performing and low-performing students[13,15]. For this study, the item discrimination is defined as the average correlation coefficient between the item score (correct or incorrect) and the overall score for each student. Note that item score is a dichotomous variable, and thus a point-biserial (simplified Pearson's) coefficient is utilized in its calculation. Many concept-inventory researchers suggest that each item should have a discrimination score above 0.2[13,21].

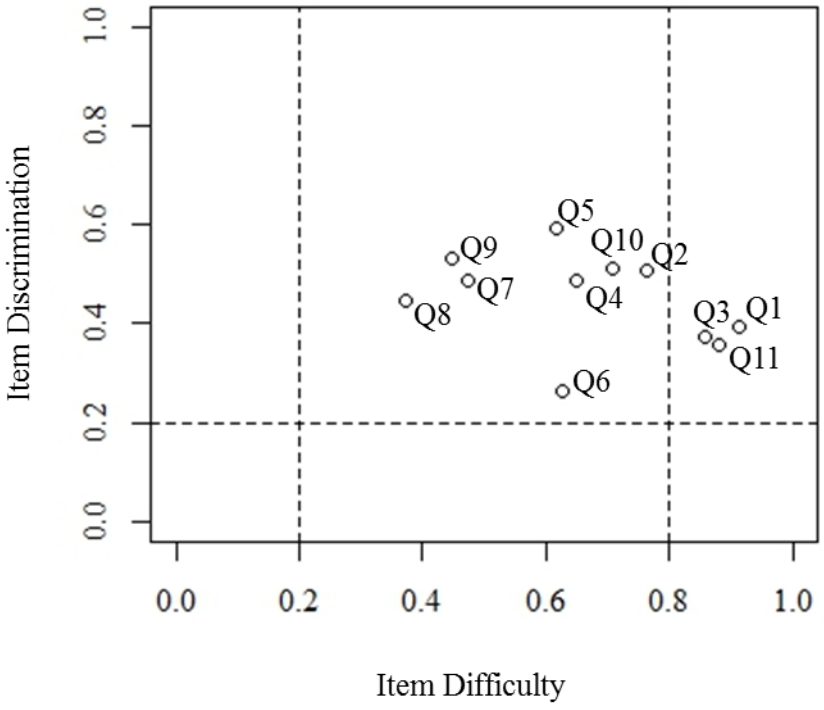
Figure 2 illustrates the fact that the item discrimination of Question 8 resided below the suggested 0.2 threshold (the dashed line on the graph) on the pre-test but increased to over 0.44 on the post-test. Question 8 of the aDCI is shown in Figure 3, and the DCI authors hypothesized that the main misconception associated with this problem is that tension is equivalent to weight[5]. While the item discrimination for Question 8 rose above the 0.2 suggested value on the post-test, it had the lowest proportion of students answering it correctly out of any of the questions included in the test. This may suggest that the tension-weight-equivalence misconception should be addressed more effectively in the curriculum.

Future Work Regarding aDCI Development

Based on the initial psychometric results presented above and those published by other researchers, it may be valuable for the Freeform researchers and the Purdue dynamics instructors to revisit the questions selected for the aDCI. It appears that for students at Purdue, three questions are overly easy and one question has an inadvisable lower alpha-with-item-deleted. Also, the aDCI could be expanded to include more questions. The increase in the number of questions would likely improve the internal reliability and validity of the aDCI, but allowances must be made in order to keep the assessment manageable for students to complete as part of their final exam.



(a)



(b)

Figure 2. Item difficulty and discrimination metrics for the (a) pre- and (b) post-test of the abbreviated DCI show that most questions adequately discriminate between low and high performing students and most questions were of appropriate difficulty. The dashed lines are commonly suggested thresholds for the metrics.

Question 8

Both systems shown have massless and frictionless pulleys. On the left, a 10 N weight and a 50 N weight are connected by an inextensible rope. On the right, a constant 50 N force pulls on the rope. Which of the following statements is true immediately after unlocking the pulleys?

(a) In both cases, the acceleration of the 10 N blocks will be equal to zero.

(b) The 10 N block on the left will have the larger upward acceleration.

(c) The 10 N block on the right will have the larger upward acceleration.

(d) The tension in the rope on the left system is 40 N.

(e) In both cases, the 10 N block will have the same upward acceleration.

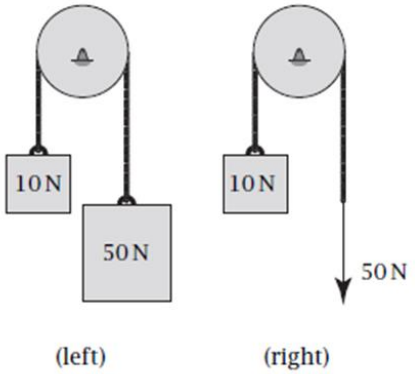


Figure 3. Question 8 of the aDCI had a lower discrimination score lower than the suggested threshold for the pre-test and the lowest correct-response rate on the post-test, suggesting the concept should receive more attention in the curriculum (adapted from Gray et al.[3]).

Using the aDCI to Preliminarily Investigate Instructor Influence on Student Performance

The development of the aDCI provided researchers with a tool for assessing students' conceptual knowledge independently of problem-solving skill. The ability to evaluate these two cognitive areas independently is important given that the Freeform framework specifically emphasizes both problem solving and conceptual understanding. Therefore, in order to answer complex research questions, such as how an instructor influences student learning within the Freeform environment, assessment tools capturing both problem-solving skills and conceptual understanding are required. This next section will detail a study involving student aDCI scores, exam scores, and final grades from multiple sections of a dynamics class utilizing the Freeform framework. These metrics are analyzed as part of a preliminary investigation into the influence of instructor differences related to student performance.

Preliminary Study: Methods

For this preliminary, cross-sectional study regarding instructor influence, problem-solving and conceptual-knowledge assessment scores were compared across three sections of dynamics during the Spring 2015 semester. All three instructors were subject-matter experts in dynamics, veteran instructors, and had taught dynamics within the Freeform environment previously. Each section was evaluated for equivalence at the beginning of the semester using multiple student-performance metrics, and the sections were again evaluated for differences at the end of the semester. Post-intervention metrics included both the aDCI scores and the midterm and final exam scores, which largely tested problem-solving abilities.

Access to the gradebooks from all instructors allowed for a common grading scheme to be developed with homeworks and quizzes comprising 20% and 5% of a student's final grade, respectively. Exam scores constituted the remaining 75% of the grade. If a student's average

score on the three midterm exams was higher than their final exam score, the midterm-exam average counted as 50% of the grade and the final was worth 25%. If a student had a higher final exam score than midterm average, the final was worth 50% and the midterm average was worth 25%. As mentioned previously, for this study the dependency of the final exam and final course grade on the aDCI (because the aDCI was part of the final exam) was eliminated by removing the aDCI component from all grade calculations.

The three dynamics sections (n =150, 107, and 104, respectively) during Spring 2015 provided convenient samples for this preliminary work because most assessments were common across sections (quizzes being the exception). The goal was not to discriminate 'good' instructors from 'bad'; instead, the hypothesis was that the Freeform system may actually minimize instructional differences across sections—thereby leveling the student experience—because of the scaffolding and support resources it provides. The available evidence from Spring 2015 was examined to determine whether this initial hypothesis was supported by the data.

Evaluation of Incoming Student-Cohort Equivalence Across Sections

To investigate the equivalence of the students enrolled in each section at the start of the semester, a series of ANOVAs (Welch or Kruskal-Wallis depending on the data type) with a Type I error significance level (α) of 0.05 were used to confirm that there were no statistically significant differences in in-coming GPA, statics grade, aDCI pre-scores, or total college credits accumulated at the time of enrollment. The analysis included credits accumulated because the number of credits accumulated affects registration priority, and higher-level students could choose a section at a more desirable time. One of the sections started at 8:30 AM, which is often viewed unfavorably by students. Statics was included because it is the only formal prerequisite for the course, and the aDCI pre-scores specifically evaluated prior conceptual knowledge.

The results of the ANOVAs found no evidence of differences with regard to averages (or average ranks for ordinal data) of in-coming GPA (p = 0.46), aDCI pre scores (p = 0.89), credits accumulated (p = 0.18), and statics grade (p = 0.75). These factors are likely the most important indicators of between-section differences that would bias our estimates of instructor effects. However, we acknowledge that other observable and unobservable factors between sections may play a role in any between-section differences. The lack of evidence for differences between sections based on in-coming GPA, aDCI score, credits accumulated, and statics grade does provide strong support for any differences in end-of-semester metrics being related to the quality of the Freeform implementation and other in-semester factors rather than pre-existing differences.

Evaluation of End-of-Semester Equivalence Across Sections

End-of-semester performance metrics were also evaluated for differences between sections. This comparison included the aDCI post-test score, aDCI normalized gain, the scores of exams 1-3 (the midterm exams), final exam scores (with the aDCI dependency removed), and the final grade percentage (with the aDCI dependency removed). The aDCI scores served as the assessment of conceptual knowledge while the exam scores represented primarily problem-solving skills.

The ANOVAs for aDCI post-test scores (p = 0.34), aDCI normalized gains (p = 0.27), exam 1 (p = 0.28), exam 2 (p = 0.08), exam 3 (p = 0.22), and final grade % (p = 0.15) established that the averages (or the ranks of the averages for ordinal data) were statistically equivalent between class sections. However, the final exam averages were statistically different across sections (p = 0.0074) with a Games and Howell[22] post-hoc analysis indicating that section 3 (average = 94%) had a significantly higher mean than sections 1 and 2 (average = 91% for both). The equivalence of all end-of-semester metrics except one suggests that students are able to achieve equivalent academic outcomes in dynamics regardless of differences in instructor pedagogical choices or differing in-class environments across sections.

The high averages on the final exam (which only comprised of scores on traditional, problem-solving questions after the aDCI scores were extracted) indicate that the students performed exceptionally well on problem-solving tasks. The mean scores of the aDCI for all sections was 66% (SD = 20%), signifying that students still struggled with conceptual understanding more than traditional problem solving. While an average performance of 66% is by no means ideal, it should be noted that this average is quite high when compared to performance averages on concept inventories in similar studies[3,15].

While the student performance was almost-entirely equivalent across sections in the study detailed above, there could be many reasons for this. For example, students may have performed equivalently across sections in a more traditional learning environment. Alternatively, the sectional equivalence could be attributed to all of the sections having veteran instructors with prior experience teaching in the Freeform environment. It would be premature to say that the cause for this equality was solely due to the fact that these three courses were taught within the Freeform environment. However, this analysis provides preliminary support for our hypothesis regarding instructor impact. Specifically, we hypothesize that the resources and support of the Freeform system may reduce the impact of instructional differences and provide similar experiences and equal opportunities for success to all students (who have diverse learning preferences and approaches) in all sections of course. An experimental design comparing students in a Freeform environment to students in a traditional environment is not possible at Purdue University because the dynamics instructors have committed fully to Freeform. However, opportunities for an experimental study may be possible in the future, as other institutions experiment with (or gradually adopt) Freeform. In the event of such an opportunity, this future research undoubtedly will be aided by the existence and use of the aDCI.

Conclusions

This paper highlights the development of an abbreviated version of the pre-existing Dynamics Concept Inventory and its subsequent application in a unique learning environment, known as Freeform. Freeform is a course system that combines active, blended, and collaborative learning pedagogies. Based on data obtained from one preliminary study of a dynamics class with three sections at Purdue University, the aDCI was shown to be sufficiently reliable and valid to include as the conceptual portion of the final exam. The implementation of the aDCI allowed researchers to assess students' conceptual understanding of dynamics on an aggregated (class or section) basis. These results lay the groundwork for using the aDCI for both cross-sectional and

longitudinal studies and as part of the examination of complex research questions. For example, in this paper, the aDCI helped evaluate the effect of the instructor on student performance within the Freeform environment.

Six out of seven post-intervention performance metrics—including the aDCI post-test scores, aDCI gains, midterm exam scores (which primarily assess problem-solving skills), and final grades—across three class sections taught by three different instructors, revealed no significant differences in student performance. While this initial study was not designed to produce generalizable results applicable to all instances of Freeform implementation, the evidence so far is consistent with the hypothesis that the strong scaffolding of the lecturebook, the immense library of online resources, and the online and physical community fostered by the Freeform framework may play a role in smoothing out instructor and student-performance differences across sections of the course. The data presented here are suggestive, and certainly not conclusive, but they indicate promise for the aDCI as a research tool for use in complex and in-depth investigations into the aspects and implications of the Freeform system.

Acknowledgements

The authors would like to thank the developers of the DCI and encourage readers to learn more about the DCI at http://www.esm.psu.edu/dci/. The material detailed in the present manuscript is based upon work supported by the National Science Foundation under Grant No. DUE-1525671. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Note to Readers

The research team for this study includes graduate research assistants, faculty members, and professionals from Purdue University's School of Engineering Education, School of Mechanical Engineering (including the Freeform creators), Center for Instructional Excellence, and the Office of Institutional Research, Assessment, & Effectiveness. Concerted efforts have been taken to analyze and report results in an unbiased and objective manner.

References

1. Froyd JE, Ohland MW. Integrated Engineering Curricula. *J Eng Educ*. 2005. **94**(1):147-164.
2. Rhoads JF, Nauman E, Holloway B, Krousgrill C. *The Purdue Mechanics Freeform Classroom : A New Approach to Engineering Mechanics Education*. ASEE Annual Conference and Exposition, Indianapolis, IN. 2014.
3. Gray GL, Costanzo F, Evans D, Cornwell P, Self B, Lane JL. *The Dynamics Concept Inventory Assessment Test : A Progress Report and Some Results Introduction*. ASEE Annual Conference and Exposition, Portland, OR. 2005.
4. Evans DL, Gray GL, Krause S, et al. *Progress on Concept Inventory Assessment Tools*. ASEE/IEEE Frontiers in Education Conference, Boulder, CO. 2003.
5. Gray GL, Evans D, Cornwell PJ, Costanzo F, Self B. *Toward a Nationwide Dynamics Concept Inventory Assessment Test*. ASEE Annual Conference and Exposition, Nashville, TN. 2003.
6. Smith MK, Wood WB, Knight JK. The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics. *CBE - Life Sci Educ*. 2008. **7**(4):422-430.
7. Henderson C. Common Concerns about the Force Concept Inventory. *Phys Teach*. 2002. **40**(9):542-547.
8. Donnellan MB, Oswald FL, Baird BM, Lucas RE. The Mini-IPIP Scales: Tiny-Yet-Effective Measures of

the Big Five Factors of Personality. *Psychol Assess*. 2006. **18**(2):192-203.
9. Gosling SD, Rentfrow PJ, Swann WB. A Very Brief Measure of the Big-Five Personality Domains. *J Res Pers*. 2003. **37**(6):504-528.
10. Han J, Bao L, Chen L, et al. Dividing the Force Concept Inventory into Two Equivalent Half-Length Tests. *Phys Rev Spec Top - Phys Educ Res*. 2015. **11**(1):010112.
11. Steif PS. *Initial Data from a Statics Concept Inventory*. ASEE Annual Conference and Exposition, Salt Lake City, UT. 2004.
12. Jorion N, Self B, James K, Schroeder L, Dibello L, Pellegrino J. *Classical Test Theory Analysis of the Dynamics Concept Inventory*. ASEE PSW Section Conference, Riverside, CA. 2013.
13. Engelhardt P V. An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Get Started PER*. 2009. 1-40. http://www.per-central.org/items/detail.cfm?ID=8807.
14. Jorion N, Gane BD, DiBello L V, Pellegrino JW. *Developing and Validating a Concept Inventory*. ASEE Annual Conference & Exposition, Seattle, WA. 2015.
15. Jorion N, Gane BD, James K, Schroeder L, DiBello L V., Pellegrino JW. An Analytic Framework for Evaluating the Validity of Concept Inventory Claims. *J Eng Educ*. 2015. **104**(4):454-496.
16. Ates S, Cataloglu E. The Effects of Students' Cognitive Styles on Conceptual Understandings and Problem-Solving Skills in Introductory Mechanics. *Res Sci Technol Educ*. 2007. **25**(2):167-178.
17. Malone KL. Correlations Among Knowledge Structures, Force Concept Inventory, and Problem-Solving Behaviors. *Phys Rev Spec Top Educ Res*. 2008. **4**(2):020107.
18. Hake RR. Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *Am J Phys*. 1998. **66**(1):64-74.
19. Nieminen P, Savinainen A, Viiri J. Relations between Representational Consistency, Conceptual Understanding of the Force Concept, and Scientific Reasoning. *Phys Rev Spec Top Educ Res*. 2012. **8**(1):010123.
20. Steif PS, Dantzler JA. A Statics Concept Inventory: Development and Psychometric Analysis. *J Eng Educ*. 2005. **94**(4):363-371.
21. Allen K, Reed-Rhoads T, Terry RA, Murphy TJ, Stone AD. Coefficient Alpha: An Engineer's Interpretation of Test Reliability. *J Eng Educ*. 2008. **97**(1):87-94.
22. Games PA, Howell JF. Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study. *J Educ Stat*. 1976. **1**(2):113-125.