International Journal of Engineering Education Vol. 35, No. 2, pp. 491–509, 2019
Printed in Great Britain

0949-149X/91 $3.00+0.00

A Validation and Differential Item Functioning (DIF) Study of an Abbreviated Dynamics Concept Inventory*

NICK A. STITES
School of Engineering Education, MEERCat Purdue: The Mechanical Engineering Education Research Center at Purdue University, Purdue University, West Lafayette, IN, USA. E-mail: nstites@purdue.edu

KERRIE A. DOUGLAS
School of Engineering Education, Purdue University, West Lafayette, IN, USA.

DAVID EVENHOUSE, EDWARD BERGER and JENNIFER DEBOER
School of Engineering Education, School of Mechanical Engineering, MEERCat Purdue: The Mechanical Engineering Education Research Center at Purdue University, Purdue University, West Lafayette, IN, USA.

JEFFREY F. RHOADS
School of Mechanical Engineering, MEERCat Purdue: The Mechanical Engineering Education Research Center at Purdue University, Purdue University, West Lafayette, IN, USA.

Concept inventories (CIs) have become popular assessment tools in science, technology, engineering, and mathematics education. Some researchers use CI scores when looking at differences in conceptual understanding or learning gains across demographic groups, but very few CIs have been evaluated for measurement bias or other aspects that threaten the fair assessment of learners. The most common psychometric evaluation models are shaped primarily by the majority demographic group, so these models can hide biases in the assessment against minority groups. The purpose of this study was to evaluate the extent to which the validity, reliability, and fairness evidence supports the use of the total score on a 12-item Abbreviated Dynamics Concept Inventory (aDCI) as a measure of a student's overall conceptual understanding of dynamics. Because of the strong relationship between the aDCI and the Force Concept Inventory, which has previously been shown to include item-level gender biases, we examined threats to fair measurement across gender scores of the aDCI. We employed an argument-based validation approach which tested: (1) the fit of a single-factor latent structure for the aDCI scores via a confirmatory factor analysis (CFA), (2) the difficulty and discrimination of each item using item response theory, (3) the correlation between the aDCI scores and similar measures of conceptual understanding, and (4) the differential item functioning of the aDCI items across gender groups via a multiple-group CFA. We found that one item had face-level construct validity concerns and two others were slightly biased against women. Possible sources of gender bias included the question's content and context. Our results suggest that the interpretation of a student's total aDCI score should consider the differential item functioning of two items across gender and the construct-alignment concerns of a third item. This work highlights the importance and challenge of designing inclusive assessments and validating them with fair psychometric models.

Keywords: concept inventory; validity; reliability; fairness; gender; engineering education; assessment bias

1. Introduction

Research suggests that a student's conceptual understanding of fundamental engineering topics directly relates to their ability to solve problems and apply existing knowledge to new and novel situations [1–4]. Concept inventories (CIs) are increasingly-popular instruments for assessing students' conceptual understanding, as well as their misconceptions, within a particular domain (such as statics, dynamics, or thermodynamics) [5]. The interest in CIs in engineering increased significantly in the early 2000s, potentially driven by a transition of ABET accreditation guidelines to a focus on program outcomes [6]. Currently, the development and assessment of conceptual understanding is still a large endeavor; a search of the US National Science Foundation awards found over $7 million in active awards with the phrase "concept inventory" in the proposal abstract alone. CIs are commonly used to evaluate pedagogical innovations [7–9], and they have also been used to better understand how students develop conceptual understanding [10]. Yet, despite the investment in and positive outcomes associated with CI use, research on the quality and fair use of these assessment instruments is generally incomplete [5].

Researchers have used many different types of evidence to validate the use of CIs, with varying degrees of quality [11]. Because validity pertains to justifying specific interpretations and uses of assessment scores, evidence must be collected to test the plausibility of the desired claims made from the scores [12]. Generally speaking, developers and

* Accepted 30 October 2018.

users of CIs have similar desired inferences from the CI scores—the students' conceptual understanding of a specific topic. Therefore, in response to the need for more consistency, researchers have begun to develop guidelines to aid those interested in developing or using CIs for their research. Streveler et al. [11] demonstrated how the Assessment Triangle can be applied to the development and testing of CIs, where evidence to support the interpretation of CI scores was empirically gathered through studies of item difficulty and discrimination. The Assessment Triangle provides a framework for assessment development that ensures the alignment between cognitive theory, observing the students' assessment responses, and interpreting the responses [13]. Focusing on the interpretation corner, Jorion et al. [5] suggested a framework to evaluate the plausibility of three common claims made from CI results: (1) students' overall conceptual understanding, (2) students' understanding of specific concepts, and (3) students' propensity for misconceptions. While these frameworks are helpful for developing and evaluating CIs, the examples do not consider use among diverse learners. According to the *Standards for Educational and Psychological Testing*, high-quality assessments are based on evidence of reliability, validity, and fairness [14]. Fair assessment has received relatively little attention in engineering education with few examples of what is meant by 'fair' and how to measure it. This work provides one example of how to operationalize and measure fairness and, to our knowledge, represents the first psychometric analysis of an engineering CI to consider fairness.

Psychometric models used in the validation of assessment instruments, such as CIs, are based on statistics for which the responses of the demographic majority group will have the most power in shaping the model. Given that only approximately 20% of U.S. engineering students are women [15], any psychometric model from that sample is essentially normed on the responses from men. To examine how the items perform for minority students, researchers need to purposefully examine measurement models for minority groups. In a recent review of assessment development articles published in engineering education journals, only one article considered potential bias in the assessment items themselves [16]. Yet, recent research in engineering education assessment validation found items that had acceptable fit for the whole student group also contained item-level bias against English-language learners [17]. An acceptable psychometric model fit for the whole group does not guarantee that those same items are fair for all students. The evaluation of test items is a prerequisite to the evaluation of the learners who responded to those items. Score differences found between groups cannot be justifiably interpreted as true score differences unless bias in the measurement model has been ruled out. It is simply unknown whether assessment questions or answer choices are understood in the same way by diverse groups of students, unless the evidence is specifically sought. Therefore, the evaluation of CIs in engineering for fairness across all minority groups, including those based on gender, is a prerequisite for the fair use of the students' scores to make decisions of personal consequence to the students.

1.1 Purpose of the study and research questions

The overall purpose of this research is to report the development and initial validation studies of a shortened form of the Dynamics Concept Inventory [18], which is named the Abbreviated Dynamics Concept Inventory (aDCI). While the DCI is an established instrument used in engineering education research (e.g., [19, 20]), a shortened version would enable instructors and researchers to assess students' conceptual understanding of dynamics in less time [21]. The research question that guides this work is: to what extent does the validity, reliability, and fairness evidence support the use of aDCI scores as a measure of students' overall conceptual understanding of dynamics? Regarding the fairness of the aDCI, we focus on gender fairness in this paper because the aDCI is closely related to the Force Concept Inventory (FCI) that is used in physics education, and the FCI has been shown to include item-level gender bias [22]. Additionally, previous research has indicated a statistically-significant gender gap in the students' total scores on the aDCI [23].

In accordance with Messick's [24] description of validation research as hypothesis testing and Kane's [25] argument-based approach to validation, we investigate the overarching research question by testing the following hypotheses:

If a student's total score on the aDCI can be interpreted as a measure of their overall conceptual understanding of dynamics, then:

- Hypothesis 1. A single-factor latent structure would effectively model the shared variance of the aDCI items;
- Hypothesis 2. The aDCI items would be appropriately difficult and able to discriminate between students with high and low overall conceptual understanding of dynamics;
- Hypothesis 3. The aDCI total score would be correlated to similar measures of overall conceptual understanding of dynamics;
- Hypothesis 4. The aDCI items would function similarly for students of equal ability regard-

Table 1. An overview of the analytical methods used in this study and the purpose of each of the hypotheses

Hypothesis	Analytical Method	Purpose
1. Single Factor Latent Structure	Confirmatory Factor Analysis (CFA)	Determine if all items of the aDCI serve as indicators of a single latent construct that is assumed to be a student's overall conceptual understanding of dynamics.
2. Appropriate Difficulty and Discrimination	Item Response Theory (IRT)	Investigate how well the difficulties of the aDCI items match the latent abilities of the students and how well the items differentiate the higher- and lower-performing students.
3. Correlated to Similar Measures	Correlation	Evaluate the relative relationships between similar measures of students' overall conceptual understanding of dynamics.
4. Measurement Invariance Across Groups	Multiple-Group Confirmatory Factor Analysis (MG-CFA)	Determine if the aDCI functions the same for men and women; i.e., evaluate the aDCI for gender bias.

less of their background or socialization, including gender.

The purpose of each hypothesis and the analytical methods used to investigate the hypotheses are summarized in Table 1. The structure of this study will follow the order of the hypotheses. The results of Hypotheses 1–3 provide information regarding the reliability and construct validity of the aDCI scores, and Hypothesis 4 targets fairness.

2. Literature Review

2.1 Reliability, validity, and fairness

The cornerstones of high-quality assessments reside in the evidence of reliability, validity, and fairness [14]. Reliability refers to the degree of consistency both internal to the assessment and of the scores for multiple administrations of the assessment [16]. Validation is the process of identifying multiple sources of relevant evidence to make a judgement about the appropriateness of using a given assessment for a specific purpose [26]. Thus, validity refers to the evidence and rationale for claiming an assessment score can be interpreted and used as intended—as a measure of the learners' knowledge, skill, or conceptual understanding [12, 16]. Of the three cornerstones of high-quality assessments, validity is overarching. Validity depends on the evidence of reliability and fairness; for an assessment to have a valid use, it must first demonstrate reliability and fairness in assessing learners.

There is no one set of procedures for validation because the validation process depends on the specific interpretation and purpose of the assessment [27]. In order to holistically evaluate the use of an assessment, one would clearly articulate the chain of reasoning involved in determining what evidence to test [27]. In the case of concept inventories used in physics or engineering education, after the assessment is administered, validity testing would begin with "If this assessment score truly measures the students' conceptual understanding, then what else has to be true so that the reliability, validity, and fairness evidence supports this argument?"

While most engineering education researchers are at least aware of the terms "reliability" and "validity" in educational assessment, fairness is less understood. Fairness was recently raised to the same level of importance as validity in the *Standards for Educational and Psychological Testing* in order to emphasize how crucial evidence of fairness is for ethical education assessment [14]. The term itself, *fairness*, does not have one specific technical meaning, as it has been used in a variety of ways in educational assessment [14]. The *Standards* identify common views of fairness to include equitable treatment during the testing process, lack of measurement bias, access to content assessed, and valid interpretations of individual test scores. Fair and valid interpretations of test scores can depend on, among other factors, the content assessed and the context of the questions [14, 28]. Measurement bias and valid interpretation of individual test scores are the most pertinent views of fairness for this work because they are partially dependent on item-level bias that can cause differential item functioning (DIF) across student groups, which is what we investigate in Hypothesis 4. Researchers have previously found item-level gender bias in physics CIs (e.g., [22, 29, 30]) and physics (mechanics) is closely related to dynamics. Therefore, we investigate threats to the gender fairness of the aDCI stemming from the psychometric models of evaluation and from the content and context of the questions.

2.2 Sources of gender bias in CIs

Because of the minimal research on the fairness of engineering CIs, we looked to the literature from physics education research for information regarding possible sources of gender bias in CIs. Madsen, McKagan, and Sayre [31] reviewed literature on the gender gap of physics concept inventories, and they identified six categories of factors that had evidence of a demonstrated impact on the gender gap: background and preparation (e.g., high school back-

Table 2. A gender-orientation framework for evaluating items on the aDCI for gender bias [34]

Criteria	Masculine Orientation	Feminine Orientation	Allegedly Neutral Orientation	Gender-Inclusive Orientation
Language	Uses he, him, his	Uses she, her, hers	Uses they, them, their Uses role (e.g., a sprinter . . .)	Uses the name of a person Uses "you"
Portrayal of stereotypes	Men in active roles, women in passive roles	Women in active roles, men in passive roles	Genderless people in active roles (e.g., a scientist . . .)	Both men and women in active and passive roles
Appeal to background experiences	Relevant to stereotyped experiences of men	Relevant to stereotyped experiences of women	Not relevant to human experiences	Relevant to men and women equally
Context	Decontextualized, abstract	Human, social	Concrete setting	Human, social, environmental

Note. Rennie and Parker used the terms *male* and *female* rather than the terms *masculine*, *feminine*, *men*, and *women* (as shown). Rennie and Parker included the word "allegedly" to the Neutral Orientation category because their research indicated that students assume plural pronouns and genderless people refer to men.

ground), gender gaps on other measures (e.g., average exam scores), differences in personal beliefs and the answer a "scientist" would give, teaching method (e.g., level of interactive engagement), stereotype threat, and question wording.

Regarding question wording, the conclusion of Madsen and colleagues was largely based on McCullough's findings [32, 33] that students changed how they answered questions on the FCI when the question wording was revised to included everyday and stereotypically feminine contexts (rather than stereotypically masculine contexts of the traditional FCI). However, the way in which the context influenced the students' performance on individual questions was inconsistent, meaning the gender gap for the overall scores remained unchanged for McCullough's revised concept inventory. Nonetheless, McCullough's findings showed that changing the context of an individual question affects how men and women answer the question.

McCullogh's findings aligned with what Ding and Caballero [28] called a *context* effect. A context effect is when one group of students is more familiar with the non-essential features of a question (such as wording, language, or images), and this extra familiarity with the context causes DIF. Alternatively, Ding and Caballero posited that DIF could be caused by a *content* effect, which is when groups of students who have been exposed to different interventions, instruction, or experiences perform differently on an item. Unfamiliar content and contexts can create extra cognitive load which can affect a student's performance because the student must first infer the situation described in the problem statement before they can attempt to solve the problem [34]. Thus, content and context effects can favor certain groups based on their background and socialization, including gender.

To help instructors identify and eliminate gender bias in physics questions, Rennie and Parker [34] developed a framework, see Table 2, for assessing the gender orientation (masculine, feminine, allegedly neutral, or gender inclusive) of physics questions along four dimensions (language, portrayal of stereotypes, appeal to background experiences, and context). Later, McCullough [32, 33] used the same framework to categorize the items of the FCI. Leveraging the strong relationship between physics and dynamics, we used this framework to qualitatively evaluate the aDCI items for gender bias.

3. Background

The sophomore-level dynamics course required by many engineering majors is often challenging. It is a gateway course to the more specialized upper-division engineering courses, and, when paired with statics, it creates the problem-solving and conceptual foundation for much of the curriculum in many engineering disciplines. To be successful in dynamics, a student must understand algebra, differential equations, vector math, physics, and statics. The incorporation of so many fundamental subject areas of engineering may be a partial explanation of why students' exam scores for dynamics courses are lower than they are in statics and thermodynamics courses [35]. Many researchers have discussed the difficult aspects of dynamics (e.g., [1, 36–38]), many of which involve prerequisite material. The difficulties that the students have with the prerequisite fundamentals support the conclusion of Gray et al. [18] that "student misconceptions are not random, but are generally the result of a deficiency in their understanding of fundamental principles." Accordingly, Cornwell [39] noted that when students do not understand the fundamentals of dynamics, they struggle to identify when or why to apply a given model or solution approach.

To help instructors assess their students' conceptual understanding of the fundamental topics of

dynamics, Gray and colleagues [18] developed the Dynamics Concept Inventory (DCI). The DCI stemmed from the need to quantitatively assess the efficacy of pedagogical innovations in dynamics. Gray et al. conducted a modified Delphi process, focus groups, student interviews, (informal) instructor interviews, and pilot tests to develop the DCI. The final result was a 29-item instrument that targeted 11 of the most important and difficult concepts in dynamics [18]. Each item included five answer choices. Psychometric analyses have found that the DCI should be used for low-stakes assessment and that the total scores could be interpreted as the students' overall understanding of concepts on the DCI [5, 18]. Thus, it is plausible that a carefully-selected subset of DCI items could provide a similar measure of the students' conceptual understanding.

3.1 aDCI Development

To streamline the implementation of a dynamics CI and to save class time [21], a shortened version of the DCI (the aDCI) was developed and incorporated into the final exam of a dynamics course (Jorion et al.'s [5] suggestion of using the DCI in a low-stakes environment was not yet published). The number of conceptual questions on past final exams for this dynamics course typically ranged from 5 to 13. Therefore, the goal for the aDCI was to target as many of the important and difficult dynamics concepts as possible with fewer than 13 items.

The DCI developers did not specify which items targeted which concepts, and very limited psychometric information was available for the DCI at the time that the aDCI was developed (early 2015). Therefore, two of the co-authors of this paper (both subject-matter experts in dynamics) used their best judgement to categorize the DCI items according to conceptual content. They then chose 11 items for inclusion in the aDCI that spanned 10 of the 11 conceptual categories and a twelfth item that tested pre-requisite physics knowledge. The questions were selected based on clarity and alignment with the material taught in the dynamics course, which reflected the curriculum of most undergraduate dynamics courses and included the study of particle and rigid-body kinematics and kinetics in two and three dimensions. The twelve selected items for the aDCI and their targeted concepts are listed in Table 3.

4. Methods

4.1 Participants and data collection

The aDCI data for this study were collected from students enrolled in a sophomore-level dynamics course at a large, public, doctoral university with the highest category of research activity [40] located in the Midwest region of the United States. The dynamics course was focused on particle and rigid-body kinematics and kinetics, as well as mechanical vibration. Each year, over 500 students enrolled in the course, often in class sections of up to 120 students. The sampling frame for this study consisted of all of the students who enrolled in the course from Spring 2015–Spring 2017. Of the 1,397 students in the sampling frame, 1,351 students completed the aDCI, and 1,250 of those students agreed to participate in the research study. The aDCI was administered as part of the course's final exam, and the items were scored as correct or incorrect (1 or 0, respectively). If an item was

Table 3. Description of the concepts assessed by each item of the aDCI (using verbatim descriptions from Gray et al. [18])

aDCI Item #	Concept Description
Q1	Newton's third law dictates that the interaction forces between two objects must be equal and opposite.
Q2	Angular velocities and angular accelerations are properties of the body as a whole and can vary with time.
Q3	If the net external force on a body is not zero, then the mass center must have an acceleration and it must be in the same direction as the force.
Q4	In general, the total mechanical energy is not conserved during an impact.
Q5	An object can have (a) nonzero acceleration and zero velocity or (b) nonzero velocity and no acceleration.
Q6	The direction of the friction force on a rolling rigid body is not related in a fixed way to the direction of rolling.
Q7	The angular momentum of a rigid body involves translational and rotational components and requires using some point as a reference.
Q8	If the net external force on a body is not zero, then the mass center must have an acceleration and it must be in the same direction as the force.
Q9	The inertia of a body affects its acceleration.
Q10	A particle has acceleration when it is moving with a relative velocity on a rotating object.
Q11	Points on an object that is rolling without slip have velocities and acceleration that depend on the rolling without slip condition.
Q12	Different points on a rigid body have different velocities and accelerations, which vary continuously.

Note. Q3 and Q8 assess the same concept.

Table 4. Demographics of the sample (N = 1250)

Variable	Value	
Major[a]	81%	Mechanical Engineering
	4%	Nuclear Engineering
	5%	Agricultural Engineering
	3%	Multidisciplinary Engineering
	6%	Other
Race/Ethnicity/ International Status	60%	Domestic, White
	7%	Domestic, Asian
	5%	Domestic, URM
	23%	International
	5%	Domestic, Other
Gender	82%	Male
	18%	Female

[a] The total percentages of major does not sum to 100% because of numeric rounding.

unanswered or if multiple answers were selected (which occurred 0.31% of the time), the response was considered incorrect. These scoring methods led to a sample with no missing data.

The demographic characteristics of our sample are shown in Table 4. The institutional-research data we used conflated race, ethnicity, and international status into one variable and collected gender as a binary variable (which we acknowledge is a simplification of the gender spectrum). The proportion of women in this course is representative of many mechanical engineering courses at large research universities in the USA, including those at the university of this study.

4.2 Data analyses

4.2.1 Preprocessing data: descriptive and correlation statistics

Prior to testing the four psychometric hypotheses, the data were explored via descriptive statistics and correlations. Because of the dichotomous nature of the data (0 = incorrect, 1 = correct), the proportions of students who answered an item correctly and inter-item tetrachoric correlation coefficients were calculated [41]. The proportions provided a measure of item difficulty [42]; the tetrachoric correlations were measures of internal reliability and how related the items were to one another [41].

4.2.2 Hypothesis 1: a single-factor latent structure

This analysis used confirmatory factor analysis (CFA) to evaluate the hypothesis that the aDCI scores reflect a unidimensional latent-factor structure, i.e., conceptual understanding of dynamics. To identify the model and estimate all the factor loadings, the variance of the latent variable was constrained to be unity. A weighted least squares estimator in the *lavaan* package (version 0.5-23.109) of R (version 3.3.2) used diagonally weighted least squares to estimate the model parameters, and it used the full weight matrix to compute robust standard errors and a mean- and variance-adjusted chi-squared (χ^2) statistic. The estimator specified the model parameters that most accurately reproduced the tetrachoric correlation matrix for the sample data.

We holistically evaluated the model through the goodness of fit statistics of χ^2, comparative fit index (CFI), and root-mean-square error of approximation (RMSEA) goodness of fit statistics. We gave the statistical significance of the χ^2 test statistic minimal consideration when determining overall model fit because of its sensitivity to sample size and non-normality [43, 44]. More weight was given to the CFI and RMSEA values. As suggested by Hu and Bentler [45], we considered CFI values above 0.950 and RMSEA values below 0.050 to be indicators of good model fit.

4.2.3 Hypothesis 2: items of appropriate difficulty and discrimination

Item response theory (IRT) models the probability of a student answering an item correctly as a function of their ability level (a latent trait) and the properties of the item that are independent of the sample. Similar to CFA, IRT utilizes a single-factor model to estimate each student's latent ability, which we again assumed to represent a student's overall understanding of dynamics. We used a 3-parameter (3PL; difficulty, discrimination, and guessing) model to characterize each item of the aDCI. The proportion of lower-performing students (those with an aDCI total score of 3 or less) who answered the item correctly was used as the initial value for the guessing parameter in the IRT model. The M2 test statistic was used to evaluate model fit, using $p < 0.050$ as the significance threshold [46]. The items' difficulty values were compared to the students' ability levels with a Wright map [5] to determine if questions were too challenging or easy for our sample. Discrimination values indicated how well the item differentiated students who knew the concept and those who did not [42].

4.2.4 Hypothesis 3: correlation with similar measures of conceptual understanding

Every intermediate exam in the dynamics course in which our participants were enrolled included conceptual questions, and in aggregate, the concepts assessed by the intermediate exams reflected the concepts assessed by the aDCI. Strong correlations between a student's performance on the conceptual questions of the three intermediate exams, their total score on the aDCI, their latent factor score from the CFA (from Hypothesis 1), and their

ability score from the IRT analysis (from Hypothesis 2) would support the assumption that these data were all measures of the students' overall conceptual understanding of dynamics. The exact concepts assessed on the intermediate exams varied slightly across semesters. To be able to compare the instructor-written questions across semesters, we standardized the students' scores for each semester individually. Regarding format, the aDCI consisted of multiple-choice questions only, and the intermediate exams incorporated multiple-choice, true/false, and short-answer conceptual questions.

4.2.5 Hypothesis 4: measurement invariance across genders

If assessment items are truly measuring the intended construct and not outside factors, there should be no group level differences in item performance. Measurement invariance refers to the assumption that the measurement model is not significantly different for different demographic groups [47]. Conversely, differential item functioning (DIF) occurs when an item functions differently for different demographic groups [48]. There are multiple methods that can be used to detect DIF including multiple-group confirmatory factor analysis [49], IRT techniques [50], and non-parametric techniques like the Mantel-Haenszel method [51]. We used multiple-group confirmatory factor analysis (MG-CFA) for this study so that we could test the invariance of the relationships between the items and the latent variable (the factor loadings) and the item thresholds (the probability that a student will answer the item correctly) independently. The testing of measurement invariance with MG-CFA involves simultaneously fitting separate measurement models (with the same latent structure) to the data from men and women. Then, differences in the parameter estimates (such as factor loadings and thresholds) across the two measurement models are investigated by sequentially adding equality constraints to the parameter estimates of both models while testing for statistically significant changes in the fit of the overall model (which includes the measurement models of both men and women).

Brown referred to four levels of increasingly strict measurement invariance as: equal form, equal factor loadings, equal thresholds, and equal indicator residuals [52]. We only tested equal form, equal factor loadings, and equal thresholds because the variances of the indicator residuals were calculated values, not estimated parameters, for our data type. For testing equal form, we compared the goodness of fit statistics and factor loadings for CFA models that used data from men only, women only, and men and women simultaneously but in separate factor structures (which we labeled Model 1). To test for equal factor loadings (Model 2), we constrained the unstandardized factor loadings for each item, respectively, to be equal across gender groups. Equal factor loadings indicate that the relationships between the items and the latent factor are the same for men and women [43, 53, 54].

The testing of equal thresholds for all of the items in aggregate (Model 3) incorporated the CFA assumption that a continuous, normally-distributed variable underlies the dichotomous score for each indicator. The threshold corresponds to the z-score that bisects the distribution curve such that the areas under the curve correspond to the proportions of students answering the questions as 0 or 1. Measurement invariance at the equal-thresholds level indicates that, on average, the items are not biased against either of the gender groups [43, 53, 54].

Because the fit statistics used to judge measurement invariance indicated how well the model reproduced the variances and covariances of the sample data overall (and in aggregate), the test of equal thresholds for all of the items simultaneously could hide biased thresholds for individual items. The test for equal thresholds for individual items was a two-phase process. First, we iteratively and individually released the equality constraint for each item's threshold to determine the statistical significance ($\Delta\chi^2$ p-value) of the change in the model fit when compared to Model 3. Second, we sequentially incorporated as many of the unequal thresholds as necessary into a final MG-CFA model. The second phase used a sequential model-improvement procedure similar to that used when altering a model based on modification indices [55, p. 733]; the baseline model was updated whenever a model with a newly-released threshold constraint fit the data better than the existing baseline model. Our modification indices (values that are used to rank model changes according to how likely the changes are to improve the model fit) were the $\Delta\chi^2$ p-values from the first phase. The thresholds for the item with the lowest $\Delta\chi^2$ p-value from the initial phase were freely estimated first, and the resulting model fit was compared to that of the baseline MG-CFA model. Then, the same testing process was repeated for the item with the second-lowest $\Delta\chi^2$ p-value, then the third-lowest, and so on, identifying a new baseline each time the release of a threshold constraint resulted in a statistically significant model-fit improvement.

The same goodness of fit indices (χ^2, CFI, and RMSEA) and their thresholds used in the prior CFA were used for the measurement invariance tests. When nested models were compared, we used a χ^2 difference ($\Delta\chi^2$) test and the change in CFI to judge if the model fit changed significantly. Because the WLS estimator adjusts the test statistic

for mean and variance, a scaled $\Delta\chi^2$ test according to Satorra's method [56] was utilized. If the p-value for a scaled $\Delta\chi^2$ test was lower than 0.050, we rejected the null hypothesis of equivalent model fits. We considered a change in CFI greater than 0.010, as suggested by Cheung and Rensvold [43], indicative of significantly different model fits.

5. Results

5.1 Descriptive and correlation statistics

The proportion of students answering each of the 12 items correctly, the inter-item tetrachoric correlation coefficients, and the item-test correlation coefficients (a measure of discrimination) are shown in Table 5. The low inter-item correlation coefficients illustrated the broad, and in some cases independent, nature of the concepts assessed on the aDCI, but Q7 had particularly low inter-item correlations, which was evident in our CFA as well. The lack of groups of items with high correlations suggested that a one-factor latent structure was the most probable model.

5.2 Hypothesis 1: a single-factor latent structure

The goodness of fit statistics for the model with a single-factor latent structure are included in Table 6 (which includes results from Hypothesis 4 also). The CFI was above the threshold for considering the model a good fit (0.950), and the RMSEA was below its 0.050 threshold. The χ^2 value had an associated p-value of less than 0.050, but this was not surprising given the dependence of the χ^2 statistic on sample size [43]. Therefore, the evidence suggests that a single-factor latent model, as shown in Fig. 1, fits the data well and supports the hypothesis that all items were indicators of a single latent construct—the overall conceptual understanding of dynamics.

Additionally, all of the factor loadings for the model shown in Fig.1 were statistically significant

Table 5. Low correlation coefficients (lower diagonal with standard errors in the upper diagonal) between items of the aDCI illustrated the broad nature of the concepts assessed by the aDCI

		Correlation Coefficients											
Item #	Proportion Correct	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Q1	0.91		0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07
Q2	0.80	**0.40**		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06
Q3	0.83	0.22	**0.35**		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06
Q4	0.63	0.18	0.28	0.29		0.04	0.04	0.05	0.04	0.05	0.05	0.04	0.05
Q5	0.63	0.21	**0.38**	0.27	0.27		0.04	0.05	0.04	0.05	0.04	0.04	0.05
Q6	0.45	0.23	0.27	**0.38**	0.21	0.24		0.05	0.04	0.05	0.04	0.04	0.05
Q7	0.67	0.07	0.12	0.18	0.06	0.13	0.18		0.05	0.05	0.05	0.05	0.06
Q8	0.41	0.18	0.19	0.07	0.25	0.26	0.14	0.08		0.04	0.04	0.04	0.05
Q9	0.36	0.26	0.23	0.07	0.21	0.21	0.12	0.17	0.28		0.04	0.04	0.05
Q10	0.42	0.13	0.25	0.08	0.13	0.26	0.21	0.19	0.18	0.20		0.04	0.05
Q11	0.66	0.22	**0.36**	0.20	0.28	0.22	0.24	0.15	0.26	**0.32**	0.38		0.05
Q12	0.84	0.12	0.24	0.22	0.22	0.22	0.18	0.09	0.10	0.15	0.13	0.23	
Total Score	0.63	0.32	0.49	0.40	0.47	0.50	0.47	0.36	0.44	0.45	0.46	0.52	0.34

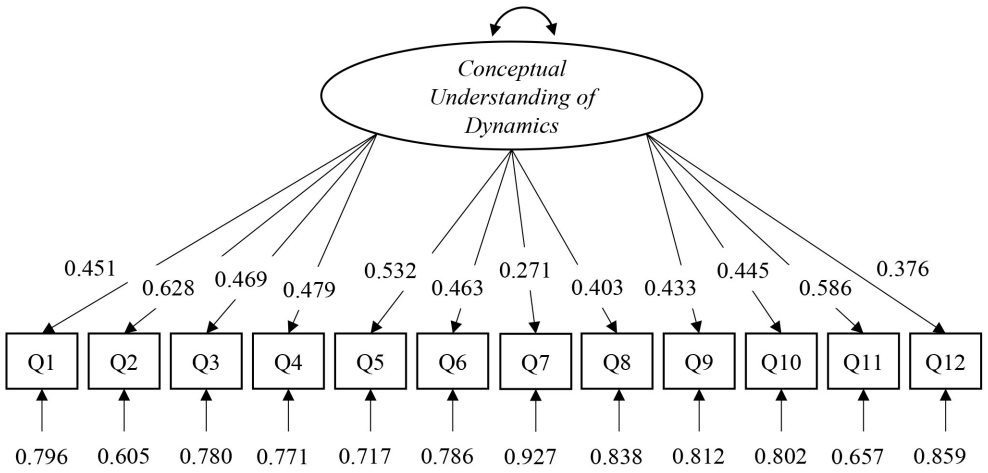

Fig. 1. The single-factor structural model fit the aggregated data from men and women well, as tested in Hypothesis 1. The numbers on the arrows from the latent construct (conceptual understanding of dynamics) to the items (Q1–Q12) represent factor loadings, which in this study are equivalent to correlation coefficients. The numbers below the items indicate the proportion of unexplained variance in each item.

Table 6. Goodness of fit and model comparison statistics for testing measurement invariance of the aDCI across men and women

		Overall Model Fit Indices						Change in Fit Indices			
Model #	**Model Description**	**df**	χ^2	χ^2 ***p*-value**	**RMSEA (90% Conf. Interval)[a]**	**CFI**	**Comparison**	**Scaled $\Delta\chi^2$**	**Scaled df**	**Scaled $\Delta\chi^2$ *p*-value**	**ΔCFI**
CFA for All Participants in Aggregate (12 items)											
–	Men and Women	54	90.08	0.002	0.028 (0.021, 0.036)	0.954	–				
Overall Measurement Invariance (11 items, Q7 removed)											
–	Men	44	70.37	0.007	0.030 (0.020, 0.039)	0.950	–				
–	Women	44	39.52	0.664	0.023 (0.000, 0.052)	0.974	–				
1	Equal Form	88	109.89	0.057	0.028 (0.018, 0.038)	0.954	–				
2	Equal Factor Loadings	99	124.91	0.040	0.023 (0.011, 0.033)	0.965	1 vs. 2	1.80	2.34	0.482	0.011
3	Equal Thresholds	109	143.74	0.014	0.025 (0.015, 0.034)	0.954	2 vs. 3	4.13	2.40	0.172	–0.011
Evaluation of Equal-Threshold Invariance for Selected Items											
4	Q3 Thresh. Est.	108	137.16	0.030	0.023 (0.000,0.052)	0.960	4 vs. 3	2.42	0.34	0.032	0.006
5	Q3, Q6 Thresh. Est.	107	131.58	0.054	0.028 (0.018,0.038)	0.965	5 vs. 4	1.66	0.27	0.044	0.005
6	Q3, Q6, Q4 Thresh. Est.	106	130.31	0.055	0.023 (0.011,0.033)	0.966	6 vs. 5	0.47	0.30	0.164	0.001

Note. $n_{men} = 1031$, $n_{women} = 219$. "Thresh. Est." indicates that an item's threshold was freely estimated across gender groups. χ^2 = chi-squared fit statistic with robust errors; df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index. Chi-squared difference tests for nested model utilized Satorra's method [56] for scaling the chi-squared statistic and df.
[a] The *p*-values for all RMSEA values listed in this table were greater than 0.990, except for the CFA for women only which was greater than 0.930.

($p < 0.001$). The factor loading for Q7, however, was nearly half that of the factor loadings for the other items. This low factor loading indicates that Q7 may be measuring a different construct than that measured by the other 11 items, and potential causes of the psychometric properties of Q7 are explored in the Discussion section.

5.3 Hypothesis 2: items of appropriate difficulty and discrimination

The M2 test statistic for the 3PL model ($M2 = 55.79$, $df = 42$, $p = 0.075$) indicated that there was no statistically significant difference between the observed data and the model-fitted data. Three conclusions can be drawn from the item characteristic curves in Fig. 2 and the parameter values in Table 7.

First, all items (except for Q7) had a positive and relatively high discrimination (maximum slope steepness) between the ability levels of –2 and 2, which was the ability range of the students in our sample. The lower discrimination value of Q7 (which is represented by the shallower slope in Fig. 2) means that Q7 did not efficiently differentiate the high- and low-ability students based on their response.

Second, Q1 and Q3 had difficulty values near negative two, which were considerably lower than most of the other questions. As shown in Fig. 3, Q1 and Q3 were most suited to differentiate students at a low ability level (near –2), and our sample had very few students with such low ability. The y-axis of Fig. 3 shows the logit transformations of the item difficulties and the students' abilities on the same scale [58]. It is preferable to have the question difficulties in the ability range with the highest density. While the power to differentiate students in our sample would have improved with higher difficulty levels for Q1 and Q3, it was expected that the students would perform well on these items because these items assessed less-challenging, prerequisite content.

Third, most of the items had non-zero guessing parameters, and many were above what would be expected for random guessing (0.20) on items with five possible answers. Therefore, the results likely indicate that students reduced the list of possible correct answers from the full set of answer choices (i.e., they eliminated poor distractors), but low-ability students still struggled to identify the correct answer from that reduced set of answers.

For example, Q12 had a guessing parameter that was considerably higher than the other items. Table 7 includes the answer distributions of lower-ability students (aDCI total score of three or lower) for all items. The answer distribution for Q12 suggests that answers A, C, and (to a lesser extent) E were poor distractors. The probability of randomly selecting the correct answer out of the two remaining choices is 50% which is close to the guessing parameter for Q12. Therefore, we are not overly concerned with the high guessing parameter of Q12 because the lack of effective distractors likely explains its high value.

Overall, these three conclusions from the IRT analysis support Hypothesis 2, except for Q7, in that the aDCI items have appropriate difficulty and discrimination for the students in our sample.

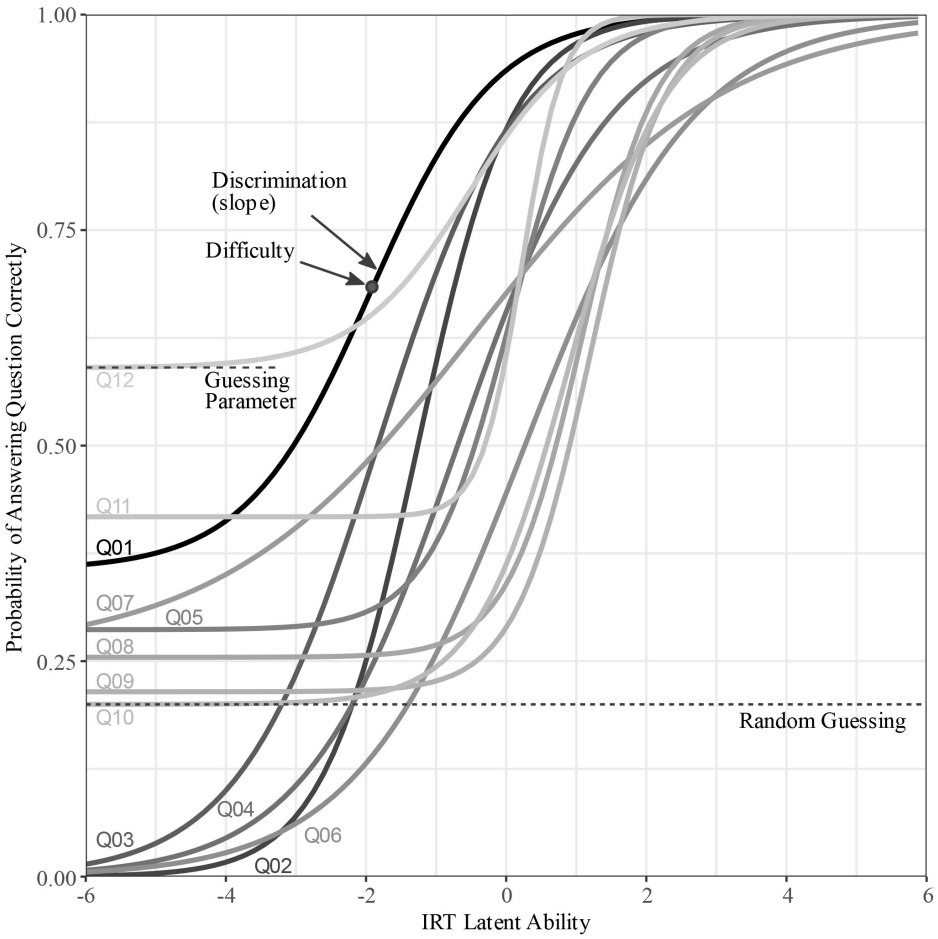

Fig. 2. A 3PL IRT model was used to fit each aDCI item to determine the discrimination, difficulty, and guessing parameters. The difficulty (inflection point of the curve and the latent ability level that bisects the sample) and discrimination (slope at the inflection point) are indicated for Q01, and the guessing parameter is illustrated for Q12. The random-guessing probability is 0.2 because all items have five answer choices.

Table 7. Summary of the IRT parameters and the answer distributions for lower-ability students who answered three or fewer questions correctly on the aDCI. Correct answers are bolded

	IRT (3PL Model)			**Answer Distribution for Lower-Ability Students**						
Question	**Discrimination**	**Difficulty**	**Guessing**	**A**	**B**	**C**	**D**	**E**	**Multiple Selected**	**None Selected**
Q1	1.13	–1.94	0.36	15	1	2	1	**33**	0	0
Q2	1.49	–1.26	0.00	25	**8**	17	1	1	0	0
Q3	1.02	–1.84	0.00	6	**16**	0	20	9	0	1
Q4	0.92	–0.69	0.00	17	7	1	20	**6**	0	1
Q5	1.72	0.05	0.29	25	6	**8**	11	2	0	0
Q6	0.83	0.27	0.00	24	16	**4**	7	1	0	0
Q7	0.55	–0.46	0.26	14	**13**	24	0	1	0	0
Q8	1.89	1.09	0.25	5	23	**7**	15	2	0	0
Q9	1.87	1.20	0.21	1	20	**1**	0	29	1	0
Q10	1.48	0.92	0.20	13	10	15	6	**8**	0	0
Q11	3.43	0.22	0.42	0	**7**	30	14	1	0	0
Q12	1.24	–0.52	0.59	2	15	3	**23**	8	0	1

5.4 Hypothesis 3: correlation with similar measures of conceptual understanding

The correlation coefficients between the students' overall performance answering conceptual questions on the three intermediate exams for the dynamics course, their factor scores (from the CFA analysis), their ability scores (from the IRT analysis), and their total aDCI score are shown in Table 8. All of the correlation coefficients were statistically significant ($p < 0.050$), and their magni-

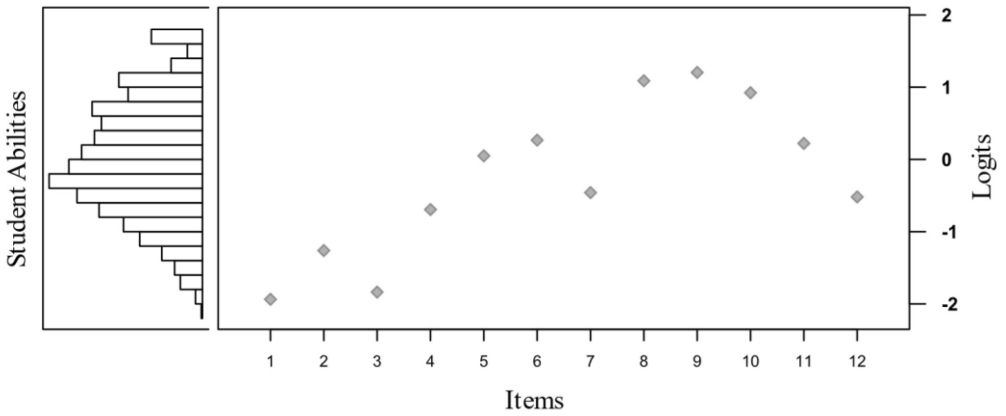

Fig. 3. An item-person map (or Wright map) of the ability scores of the participants and the difficulty values for the 12 items of the aDCI. A logit (vertical axis) is the natural log transformation of (1) the odds ratio for answering an item correctly, or (2) the ratio of a students ability divided by one minus their ability [58]. An item for which a student, in general, would have a 50% chance of answering correctly would have a logit of zero, and an average-performing student with an ability of zero would have a logit of zero.

Table 8. Strong correlations (coefficients in the lower diagonal with standard errors in the upper diagonal) between different measures of the students' conceptual understanding suggest that they all may be measuring the same construct

	Exam Questions	CFA Scores	IRT Abilities	aDCI Total Score
Exam Questions		0.03	0.03	0.03
CFA Scores	0.46		0.03	0.03
IRT Abilities	0.44	0.99		0.03
aDCI Total Score	0.46	0.99	0.97	

tudes with the instructor-written questions correspond with a medium effect size [59], suggesting that they measured similar (if not the same) constructs as proposed in Hypothesis 3.

Fig. 4 shows that the relationship between the students' total aDCI scores and the CFA factor scores (which were highly correlated with the IRT ability levels) was linear and highly correlated. This relationship allows for the aDCI total scores to be used as a proxy measurement of the students' overall conceptual understanding of dynamics without having to conduct a CFA or IRT analysis.

5.5 Hypothesis 4: measurement invariance across genders

The tetrachoric correlation coefficients used in the MG-CFA are shown in Table 9 (Q7 was not included in this analysis because of poor psychometrics in Hypotheses 1–3). On average, the correlation coefficients for women exceeded those for men, foreshadowing that a single-factor latent structure will fit the data from women better than it will for the aggregated or men-only data.

Table 6 shows the results from the MG-CFA that was used to investigate the invariance of the mea-

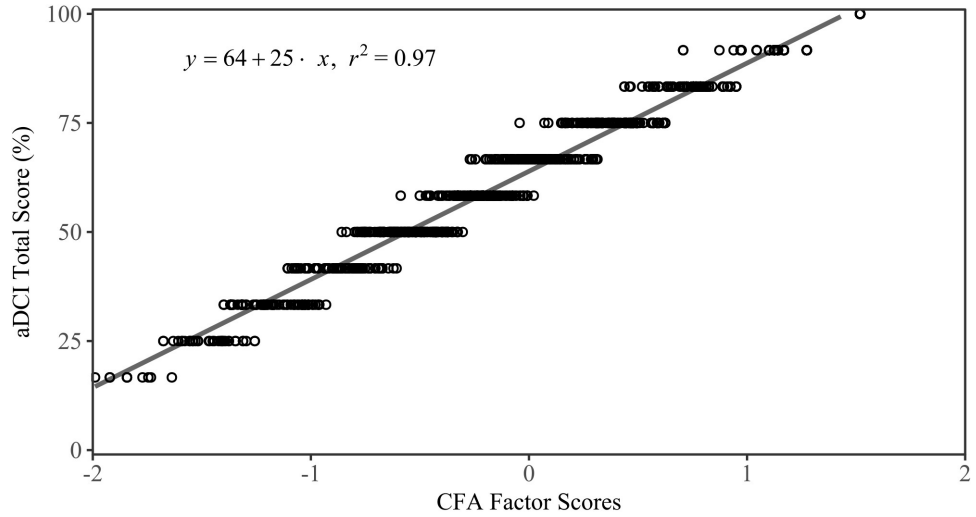

Fig. 4. The students' total aDCI scores can be used as a measure of conceptual understanding because the aDCI scores are linearly related and highly correlated with the latent factor score from the CFA.

Table 9. The correlations coefficients between the aDCI items for men (below the diagonal) and women (above the diagonal).

Item #	Q1	Q2	Q3	Q4	Q5	Q6	Q8	Q9	Q10	Q11	Q12
Q1		**0.54**	0.27	0.20	0.03	–0.08	0.15	**0.34**	–0.08	0.18	0.13
Q2	**0.32**		0.34	0.41	0.35	0.26	0.30	0.27	0.16	**0.35**	0.15
Q3	0.14	**0.32**		0.27	0.14	0.27	0.08	0.16	0.10	0.21	0.30
Q4	0.16	0.24	0.29		**0.38**	0.26	0.20	**0.31**	0.03	**0.32**	0.24
Q5	0.25	**0.37**	0.29	0.24		0.12	0.26	0.24	0.11	0.11	**0.35**
Q6	0.29	0.25	**0.38**	0.19	0.24		0.10	0.24	0.14	0.26	0.12
Q8	0.18	0.15	0.05	0.26	0.24	0.12		0.22	0.03	0.26	0.20
Q9	0.22	0.21	0.02	0.19	0.19	0.09	0.28		**0.32**	0.40	0.18
Q10	0.19	0.27	0.06	0.15	0.28	0.22	0.20	0.18		**0.39**	0.06
Q11	0.21	**0.35**	0.17	0.26	0.24	0.22	0.25	0.29	**0.38**		0.18
Q12	0.08	0.25	0.15	0.20	0.17	0.17	0.07	0.13	0.14	0.22	

Note. The correlations coefficients greater than 0.30 are bolded. The shaded cells indicate a correlation coefficient less than 0.30 for one gender with a corresponding coefficient greater than 0.30 for the other gender. $n_{men} = 1031$, $n_{women} = 219$.

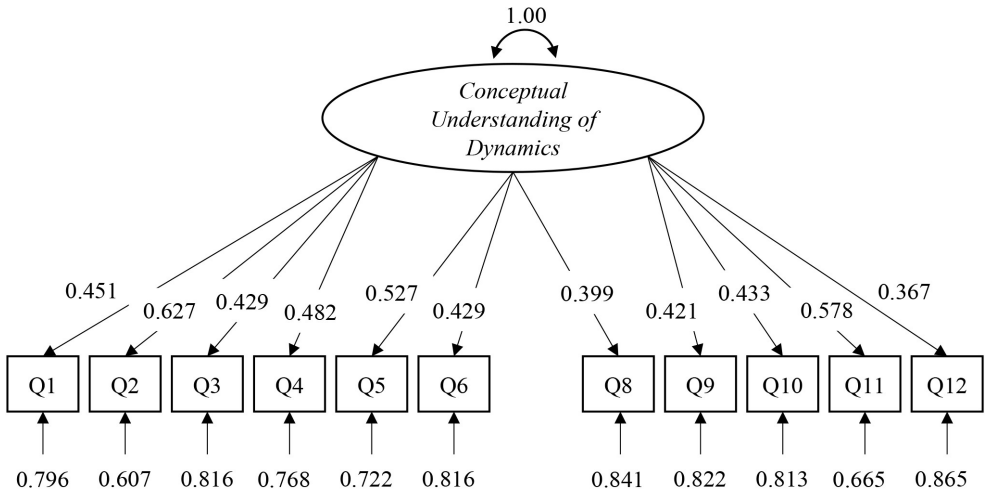

Fig. 5. The test results of Hypothesis 4 showed the aDCI scores to have invariant form, factor loadings, and thresholds across men and women. All of the factor loadings were statistically significant with p-values < 0.001.

surement models for men and women. The change in fit statistics for Models 1–3 indicated that the CFA models for men and women had equal form, equal factor loadings, and equal thresholds when considering all of the items in aggregate. For Model 1 (equal form), the significant χ^2 p-value was likely an artifact of a large sample size. For Model 2 (equal factor loadings), the positive change in CFI was likely an artifact of the CFI calculation incorporating model complexity [60,61], and Model 2 was less complex than Model 1 because of the constraints imposed on the factor loadings. Model 3 (equal thresholds—overall) had a scaled $\Delta\chi^2$ p-value that was much higher than the 0.050 criterion and a negative change in CFI (indicating a worse fit) that was only slightly outside of the recommended threshold of 0.01. Overall, the evidence suggested that the aDCI scores were measurement invariant at the threshold level when considering all of the items simultaneously.

In the first phase of our testing for equal thresholds for individual items, Q3 was identified as the item with the lowest $\Delta\chi^2$ p-value ($p = 0.032$), followed by Q6 ($p = 0.057$), Q4 ($p = 0.091$), and the rest of the items. Accordingly, Model 4 freely estimated the threshold for Q3 and was compared to Model 3. The lower χ^2 value, a $\Delta\chi^2$ p-value of less than 0.050, and the positive change in CFI indicated that Model 4 fit the data better than Model 3. The change in CFI of less than 0.010, however, illustrated only a small improvement in model fit. When considered together, the two model-difference statistics suggested that item Q3 was biased against women, but the magnitude of the bias was relatively small. The comparison of Model 5 (with the thresholds for Q3 and Q6 freely estimated) to Model 4 yielded a similar conclusion: Q6 was biased against women, but the bias was small.

The analysis continued in a similar fashion for Q4 (Models 6) and the rest of the items (not shown), but the change in fit indices suggested that none of these models statistically improved the goodness of fit when compared to Model 5. Thus, only Q3 and Q6 exhibited statistically-significant measurement bias across genders.

Table 10. The thresholds for nine of the eleven items were invariant across gender; Q3 and Q6 had unequal thresholds (bolded) that suggested potential bias against women

	Q1	Q2	Q3	Q4	Q5	Q6	Q8	Q9	Q10	Q11	Q12
Men	–1.38	–0.89	**–1.05**	–0.39	–0.39	**0.04**	0.19	0.31	0.17	–0.48	–1.05
Women	–1.38	–0.89	**–0.76**	–0.39	–0.39	**0.29**	0.19	0.31	0.17	–0.48	–1.05

In summary, the statistical evidence suggested that the aDCI scores were measurement invariant at the threshold level when considering all of the items in aggregate. At the item level, two items, Q3 and Q6, exhibited DIF with a slight bias against women. Fig. 5 summarizes the final measurement model, and Table 10 lists the thresholds for the eleven indicators.

6. Gender bias in the aDCI

The measurement invariance analysis suggested that Q3 and Q6 may be slightly biased against women, but the analysis does not tell why the items may favor men. To better understand the DIF of Q3 and Q6, we qualitatively evaluated Q3 and Q6 for content and context bias, as informed by our review of the physics education literature. We used Rennie and Parker's [34] gender-orientation framework, see Table 2, to identify possible content and context effects (as defined by Ding and Caballero [28]). We also consulted three gender-studies experts to aid in this gender-orientation analysis.

6.1 Description of the biased items

To maintain question security, we do not include the full copies of Q3 or Q6. Q3 originates from the FCI and involves a hockey puck sliding at a constant velocity across frictionless ice. The puck is kicked with a force perpendicular to the direction of its current motion, and the students are asked to identify the path of the puck after the kick. Five paths are pictorially provided as answer choices. Q6 describes a rear-wheel-drive car that is accelerating forward, and an image of a sports car is used to illustrate the scenario. The tires do not slip on the road, and students are asked to find the magnitude and direction of the friction force acting on the front tires. Five answer choices are provided, each with a symbolic equation for the magnitude of the friction force and a direction for the force ("to the left" or "to the right").

6.2 Bias in Q3

The categorization of Q3 according to Rennie and Parker's gender orientation framework identified the hockey context of Q3 as potentially appealing to the background experiences of men more than women. Previous FCI researchers have also advised that the hockey context of Q3 favors men [29, 32]. However, some of our experts suggested that the bias may be more geographical (favoring those from colder and Northern climates) than gender related. Overall, we concluded that the hockey context is a *possible* source of bias, but not a *definitive* source of bias for Q3.

Q3 of the aDCI was originally copied from the FCI (#8 in version 95 [62]), and research on the gender fairness of the FCI has not identified Q3 as having statistically-significant gender bias [22, 29, 30]. While it is true that some of the FCI studies do not have comparable samples to the engineering students of this study, at least two of the FCI studies ([22] and [33]) have samples from university-level, calculus-based physics courses that typically enroll science and engineering majors. We would expect students in these calculus-based physics classes to perform similarly on #8 of the FCI as the engineers in our sample do on Q3 of the aDCI.

Our results regarding the DIF of Q3 could differ from FCI research because of different analysis methods. We utilized MG-CFA to investigate measurement invariance across gender, but Traxler and colleagues [22] (who have published one of the most complete studies of gender bias in the FCI) used the Mantel-Haenszel and Lord's statistic (an IRT-based method). Because our results suggested that Q3 (and Q6) were only slightly biased against women, the differences in samples and methods could explain the contradictions between our results and those published for the FCI.

6.3 Bias in Q6

Q6 was categorized by some of our experts as appealing to the background experiences stereotypically more common for men than women because it may be more likely that men understand the meaning of "rear-wheel-drive," a phrase used to indicate that the rear tires provide the traction force required to accelerate the car forward. This opinion aligns with Ding and Caballero's [28] content effect because it reflects the perspective that boys in the USA are often socialized to know more about how automobiles work than girls [63]. However, some of our experts argued that this generalization about girls' relative knowledge of automobiles may not

hold true for women in engineering because women in engineering have self-selected into a masculine-oriented field that centers on understanding how systems and machines work. Even though there was disagreement regarding the gender bias related to the phrase "rear-wheel-drive," all of the experts agreed that the sports-car image used in Q6 would be stereotypically associated with men more than women. This image could be contributing to a context effect.

Q11 and Q12 also involved a rear-wheel-drive sports car, but these two questions did not display DIF. One explanation for this difference is that Q6 requires students to understand how a rear-wheel-drive car works, but Q11 and Q12 do not require this specialized knowledge. Q11 and Q12 pertain to the kinematics of a wheel that rolls without slipping, and the question prompt only uses a rear-wheel-drive car as the structure to which the wheel is attached. Furthermore, the primary image of Q11 and Q12 is that of a wheel and tire, not the sports car. Therefore, the likelihood of the sports-car image causing a context effect favoring men in Q11 and Q12 may have been less than that for Q6 because the students' focus was on the wheel (a gender-neutral image) and not the sports car.

7. Discussion

7.1 Review of purpose and results

The purpose of this study was to evaluate the extent to which aDCI scores can be used as a reliable, fair, and valid measure of undergraduate students' overall conceptual understanding of dynamics. We organized our inquiry around four hypotheses that focused on the evidence of the aDCI's latent structure, difficulty, discrimination, correlation with similar measures, and gender fairness (in terms of measurement bias). We review the evidence for each hypothesis below.

For Hypothesis 1, the results of the CFA suggested that a single-factor latent model fit the aDCI scores well. This unidimensional latent structure reflects the intentionality of the aDCI developers to select items from the DCI that assessed a broad range of topics to approximate the students' overall understanding of dynamics. The fit of the IRT model further supports the unidimensionality of the aDCI. The correlation of the students' factor scores with their performance on instructor-written conceptual questions provides evidence for the argument that the single latent factor (of the CFA or IRT models) represents the students' overall conceptual understanding of dynamics. Therefore, the evidence for Hypothesis 1 suggests that the students' total aDCI score can be interpreted as a measure of their overall conceptual understanding of dynamics.

The results of the IRT analysis used to test Hypothesis 2 indicated that most of the items on the aDCI have appropriate difficulty and discrimination values for the sample tested. Two items were identified as especially easy (they had high difficulty values), but high performance on these items was expected because the items targeted fundamental, particle-mechanics knowledge that the students likely learned in a prerequisite physics class. All of the items, except Q7, had high discrimination values (maximum slope steepness), meaning they reliably differentiated the higher-performing students from the lower-performing students.

In addition to a low discrimination value, Q7 also had low correlations with other items and low factor loading in the CFA, indicating it might be measuring a different construct than the other items on the aDCI. Q7 was one of the DCI items Jorion et al. [5] identified as having poor psychometric characteristics. Upon closer inspection, we found the wording of Q7 to be imprecise with multiple correct answers, depending on how the question was interpreted. Thus, multiple pieces of evidence suggest that the modification or replacement of Q7 could improve the utility of the aDCI, and a clarification of the question wording so that only one answer is correct may be all that is needed.

The results of testing Hypothesis 3 indicated that the aDCI total scores positively correlated with the students' performance on similar, instructor-written questions. The relationship between the aDCI total scores and the factor scores (from the CFA which were highly correlated to the IRT ability levels) was linear and had a high coefficient of determination. These two results provide evidence in support of the aDCI scores measuring one latent factor, and the latent factor can be interpreted as the students' overall conceptual understanding of dynamics. Wang and Bao [64] made a similar conclusion regarding their students' conceptual understanding of physics based on the linear relationship between the students' FCI scores and their IRT abilities.

For Hypothesis 4, the analysis of measurement invariance found the aDCI to have equal form, equal factor loadings, and equal thresholds for men and women when considering all of the items in aggregate. These results suggest that, on average, the aDCI functions similarly for men and women in measuring the students' overall conceptual understanding of dynamics. When considering all items in aggregate, the aDCI scores of men and women display: the same single-factor latent structure, the same relationships between the items and the latent factor, and the same probabilities of answering a

given question correctly. However, at the item level, the analysis identified two items, Q3 and Q6, that exhibited slight bias against women. The bias of these items indicates that when considering a man and a woman with equal overall understanding of dynamics, the man has a higher likelihood of answering Q3 and Q6 correctly than the woman. To understand why these items may favor men, we evaluated them for content and context bias.

The supporting and contradicting evidence for the possible sources of gender bias in Q3 and Q6 make it difficult to definitively say why these two items favor men. For Q3, the hockey context may disadvantage women. For Q6, the need to know how a rear-wheel-drive car works and/or the image of a sports car may differentially affect students' performance based on gender. The uncertainty in the sources of bias supports the need for further validation and fairness studies of the aDCI, DCI, and concept inventories in engineering more broadly.

7.2 Fairness implications

The investigation of DIF identified two items that favored men, but this bias was not evident in the psychometric models that used aggregated data. Based on the lack of research regarding the fairness of engineering education assessments [16], it is highly likely that many researchers would have found the psychometric evidence (from Hypotheses 1–3) satisfactory for their use of the aDCI scores as measures of the students' overall conceptual understanding. However, our results suggest that instructors and researchers must consider the gender bias in at least two of the aDCI items (Q7 was not tested for DIF) when interpreting women's total scores. Two additional incorrect responses (corresponding to the two biased items) yields an almost 17 percentage-point reduction in a student's total aDCI score. While our results indicate that the bias of Q3 and Q6 is small, it undoubtedly contributes to the gender gap in the aDCI scores that has been previously reported [23]. Thus, decisions made based on a student's overall aDCI performance, including the assignment of points toward their grade in a course, unfairly disadvantage women.

7.3 Limitations and future work

One limitation of this study is that it was conducted with students from a single institution. As Madsen et al. [31] determined, many findings from research on the gender gaps of physics CIs are not consistent across studies; thus, future research should consider how the aDCI functions at other institutions. Future work should also incorporate fairness studies for using the aDCI across other subgroups of students, including race/ethnicity, social economic status, academic major, and international status. Content and context effects could be especially relevant to English-as-a-second-language learners [17] and to students who have lived in cultures different than those present in the USA. An analysis of fairness for some of these subgroups (e.g., subgroups based on race/ethnicity) would especially benefit from more data because of their small sample sizes in engineering.

A second limitation of this study is the small number of women in the sample compared to men. This unbalanced sample could be hiding DIF that the measurement invariance analysis cannot detect because the fit of the model for the men's covariance matrix may have overshadowed a lack of fit for the women's covariance matrix. The sample sizes of men and women could be made equal by randomly subsampling from the pool of men, but the statistical power to detect DIF across the groups greatly decreases with this technique because of the small number of women in the sample. Given that women students are a small fraction of the overall student population in engineering, defining new norms for fair statistical models while maintaining sufficient power is a challenge for the engineering education research community.

A qualitative study of how women in the course experience the gender bias of the aDCI, as illustrated here, or other course assignments could help contextualize our findings. For example, do students (women or men) recognize gender bias in the course materials (including assessments), and does the content or context of these materials cause students to feel disadvantaged or uncomfortable? If so, in what ways do students articulate this discomfort, and what suggestions do they have for addressing it? A qualitative study, potentially including interviews with both women and men, would inform our understanding of students' experiences and could inspire changes to course materials to make them more inclusive and fair.

8. Conclusions

This study investigates the extent to which a student's aDCI total score can be interpreted as a reliable, valid, and fair measure of their overall conceptual understanding of dynamics. To our knowledge, this study is the first to implement an argument-based approach for the validation study of a CI and the first to investigate the fairness of an engineering CI. The results of our study suggest that aDCI scores, excluding Q7 which should be modified or replaced, for the men in our sample can be interpreted as measures of the students' overall conceptual understanding in dynamics with evi-

dence of: (1) broad content coverage and instructional relevancy, (2) appropriate interpretation of scores with regard to their underlying, single-factor latent construct, (3) appropriately difficult items that discriminate students based on their level of conceptual understanding, and (4) strong correlations between aDCI total scores and other measures of dynamics conceptual understanding. The total aDCI scores for women, however, incorporate two items with slight gender biases against women and, therefore, do not accurately reflect women's overall conceptual understanding of dynamics.

Unless further research refutes our results and supports the aDCI as a fair assessment tool for *all* students, or until the aDCI is modified to be gender inclusive and fair, we do not support its use in high-stakes testing, including its use on a final exam (as was done for our sample). Instead, we suggest that the aDCI, in its current form, be used as a low-stakes assessment instrument for measuring students' overall conceptual understanding of dynamics, and instructors and researchers should account for the DIF of Q3 and Q6 and the validity concerns of Q7 when making inferences from the aDCI scores. Alternatively, instructors could administer a shortened aDCI that excludes Q3, Q6, and Q7, knowing that the number of concepts assessed by a shortened aDCI would be less than the 12-item aDCI.

This work highlights the importance of designing inclusive assessments and validating their use with psychometric models that do not unfairly disadvantage certain subgroups of students—such as women. When assessments utilize validation studies that are dominated by one group of students, such as men, it is often unknown whether group differences in scores are artifacts of the assessment questions themselves, or truly representative of differences in the learners' understanding. Without evidence that the assessments themselves are truly fair for all engineering students, there is a very strong risk of educational inequity. Thus, more fairness studies of engineering education assessments are needed to better inform the academic community on what factors should be considered when designing an assessment that does not unfairly disadvantage students based on their background or socialization.

Acknowledgements—The authors would like to acknowledge the many students, staff, and faculty that made this work possible, including Craig Zywicki, Taylor Prebel, Charles M. Krousgrill, and David B. Nelson. Additionally, we would like to recognize Amy C. Moors, Cheryl Cooky, and Sharra Vostral for their expert insights on the possible sources of gender bias in the aDCI questions. 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.

References

1. R. Streveler, T. Litzinger, R. Miller and P. Steif, Learning conceptual knowledge in the engineering sciences: Overview and future research directions, *Journal of Engineering Education* **97**, pp. 279–294, 2008.
2. G. Hatano and K. Inagaki, Two courses of expertise, *Research and Clinical Center for Child Development Annual Report*, **6**, pp. 27–36, 1984.
3. A. F. McKenna, An investigation of adaptive expertise and transfer of design process knowledge, *Journal of Mechanical Design*, **129**(7), pp. 730–734, 2007.
4. M. G. Pandy, A. J. Petrosino, B. A. Austin and R. E. Barr, Assessing adaptive expertise in undergraduate biomechanics, *Journal of Engineering Education*, **93**(3), pp. 211–222, 2004.
5. N. Jorion, B. D. Gane, K. James, L. Schroeder, L.V. DiBello and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, *Journal of Engineering Education*, **104**(4), pp. 454–496, 2015.
6. T. Reed-Rhoads and P. K. Imbrie, Concept inventories in engineering education, *Board of Science Education Workshop on the Evidence on Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education*, Washington, DC, 2008.
7. S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt and M.P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proceedings of the National Academy of Sciences*, **111**(23), pp. 8410–8415, 2014.
8. R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *American Journal of Physics*, **66**(1), pp. 64–74, 1998.
9. D. Hestenes, M. Wells and G. Swackhamer, Force Concept Inventory, *The Physics Teacher*, **30**(3), pp. 141–158, 1992.
10. C. Henderson, Common concerns about the Force Concept Inventory, *The Physics Teacher*, **40**(9), pp. 542–547, 2002.
11. R. A. Streveler, R. L. Miller, A. I. Santiago-Roman, M. A. Nelson, M. R. Geist and B. M. Olds, Rigorous methodology for concept inventory development: Using the 'assessment triangle' to develop and test the Thermal and Transport Science Concept Inventory (TTCI), *International Journal of Engineering Education*, **27**(5), pp. 968–984, 2011.
12. S. Messick, Meaning and values in test validation: The science and ethics of assessment, *Educational Researcher*, **18**(2), pp. 5–11, 1989.
13. National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academies Press, Washington, DC, 2001.
14. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, AERA, Washington, DC, 2014.
15. National Science Board, *Science and Engineering Indicators 2018*, National Science Foundation, Alexandria, VA, 2018.
16. K. A. Douglas, A. Rynearson and S. Purzer, Reliability, validity, and fairness: A content analysis of assessment development publications in major engineering education journals, *International Journal of Engineering Education*, **32**(5A), pp. 1960–1971, 2016.
17. K. A. Douglas, T. Fernandez, S. Purzer, M. Fosmire and A. Van Epps, The Critical-Thinking Engineering Information Literacy Test (CELT): A Validation Study for Fair use Among Diverse Students, *International Journal of Engineering Education*, **34**(4), pp. 1347–1362, 2018.
18. G. L. Gray, F. Costanzo, D. Evans, P. Cornwell, B. Self and J. L. Lane, The Dynamics Concept Inventory assessment test: A progress report and some results, *ASEE Annual Conference and Exposition*, Portland, OR, 2005.
19. B. P. Self and J. M. Widmann, Demo or hands-on? A crossover study on the most effective implementation strategy for inquiry-based learning activities, *ASEE Annual Conference and Exposition*, Columbus, OH, 2017.
20. B. Coller, A glimpse into how students solve concept pro-

blems in rigid body dynamics, *ASEE Annual Conference and Exposition*, Seattle, WA, 2015.
21. N. A. Stites, D. A. Evenhouse, M. Tafur, C. M. Krousgrill, C. Zywicki, A. N. Zissimopoulos, D. B. Nelson, J. DeBoer, J. F. Rhoads and E. J. Berger, Analyzing an abbreviated Dynamics Concept Inventory and its role as an instrument for assessing emergent learning pedagogies, *ASEE Annual Conference and Exposition*, New Orleans, LA, 2016.
22. A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak and R. Lindell, Gender fairness within the Force Concept Inventory, *Physical Review Physics Education Research*, **14**(1), p. 010103, 2018.
23. T. Prebel, N. A. Stites, E. Berger, J. F. Rhoads and J. DeBoer, Work in Progress: Predictive analysis of conceptual understanding based on self-reported student engagement with resources in a blended engineering class, *Research in Engineering Education Symposium*, Bogota, Colombia, 2017.
24. S. Messick, *Validity of Test Interpretation and Use*, Educational Testing Service, Princeton, NJ, 1990.
25. M. T. Kane, An argument-based approach to validity, *Psychological Bulletin*, **112**(3), pp. 527–535, 1992.
26. K. A. Douglas and Ş. Purzer, Validity: Meaning and relevancy in assessment for engineering education research, *Journal of Engineering Education*, **104**(2), pp. 108–118, 2015.
27. R. J. Mislevy and G. D. Haertel, Implications of evidence-centered design for educational testing, *Educational Measurement: Issues and Practice*, **25**(4), pp. 6–20, 2006.
28. L. Ding and M. D. Caballero, Uncovering the hidden meaning of cross-curriculum comparison results on the Force Concept Inventory, *Physical Review Special Topics—Physics Education Research*, **10**(2), p. 020125, 2014.
29. R. D. Dietz, R. H. Pearson, M. R. Semak, C. W. Willis, N. S. Rebello, P. V. Engelhardt and C. Singh, Gender bias in the force concept inventory?, *AIP Conference Proceedings*, **1413**, pp. 171–174, 2012.
30. S. E. Osborn Popp, D. Meltzer and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, *Annual Meeting of the American Educational Research Association*, New Orleans, LA, 2001.
31. A. Madsen, S. B. McKagan and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, *Physical Review Special Topics—Physics Education Research*, **9**(2), p. 020121, 2013.
32. L. McCullough, Gender, context, and physics assessment, *Journal of International Women's Studies*, **5**(4), pp. 20–30, 2004.
33. L. McCullough, Gender differences in student responses to physics conceptual questions based on question context, *ASQ Advancing the STEM Agenda in Education, the Workplace and Society Conference*, Menomonie, WI, 2011.
34. L. J. Rennie and L. H. Parker, Assessment in physics: Further exploration of the implications of item context, *The Australian Science Teachers Journal*, **39**(4), pp. 28–32, 1993.
35. J. E. Froyd and M. W. Ohland, Integrated engineering curricula, *Journal of Engineering Education*, **94**(1), pp. 147–164, 2005.
36. J. Clement, Students' preconceptions in introductory mechanics, *American Journal of Physics*, **50**(1), pp. 66–71, 1982.
37. K. J. Shryock and J. E. Froyd, Alignment of preparation via first-year physics mechanics and calculus courses with expectations for a sophomore statics and dynamics course, *ASEE Annual Conference and Exposition*, Vancouver, BC, 2011.
38. B. Y. White, Sources of difficulty in understanding Newtonian dynamics, *Cognitive Science*, **7**, pp. 41–65, 1983.
39. P. J. Cornwell, Dynamics evolution—Chance or design, *ASEE Annual Conference and Exposition*, St. Louis, MO, 2000.
40. Indiana University Center for Postsecondary Research, The Carnegie Classification of Institutions of Higher Education, http://carnegieclassifications.iu.edu. Accessed 14 June 2018.
41. T. F. Scott, D. Schumayer and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, *Physical Review Special Topics—Physics Education Research*, **8**(2), p. 020105, 2012.
42. M. Tavakol and R. Dennick, Post-examination analysis of objective tests, *Medical Teacher*, **33**(6), pp. 447–458, 2011.
43. G. W. Cheung and R. B. Rensvold, Evaluating goodness-of-fit indexes for testing measurement invariance, *Structural Equation Modeling: A Multidisciplinary Journal*, **9**(2), pp. 233–255, 2002.
44. L. Hu and P. M. Bentler, Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification, *Psychological Methods*, **3**(4), pp. 424–453, 1998.
45. L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, **6**(1), pp. 1–55, 1999.
46. A. Maydeu-Olivares and C. García-Forero, Goodness-of-fit testing, in Penelope Peterson, Eva Baker, and Barry McGaw (eds), *International Encyclopedia of Education*, Elsevier, Oxford, UK, pp. 190–196, 2010.
47. N. Schmitt and G. Kuljanin, Measurement invariance: Review of practice and implications, *Human Resource Management Review*, **18**(4), pp. 210–222, 2008.
48. D. J. Bauer, A more general model for testing measurement invariance and differential item functioning, *Psychological Methods*, **22**(3), pp. 507–526, 2017.
49. R. J. Vandenberg and C. Lance, A Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research, *Organizational Research Methods*, **3**(1), 2000, pp. 4–70.
50. S. Stark, O. S. Chernyshenko and F. Drasgow, Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy, *Journal of Applied Psychology*, **91**(6), pp. 1292–1306, 2006.
51. A. Socha, C. E. DeMars, A. Zilberberg and H. Phan, Differential item functioning detection with the Mantel-Haenszel procedure: The effects of matching types and other factors, *International Journal of Testing*, **15**(3), pp. 193–215, 2015.
52. T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York, NY, 2015, 2nd ed.
53. W. Meredith, Measurement invariance, factor analysis and factorial invariance, *Psychometrika*, **58**(4), pp. 525–543, 1993.
54. R. B. Kline, *Principles and Practice of Structural Equation Modeling*, Guilford, New York, NY, 4th ed., 2016.
55. B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, Pearson, Boston, MA, 6th ed., 2013.
56. A. Satorra, Scaled and adjusted restricted tests in multi-sample analysis of moment structures, in R. D. H. Heijmans, D. S. G. Pollock and A. Satorra (eds), *Innovations in Multivariate Statistical Analysis*, Springer, Boston, MA, pp. 233–247, 2000.
57. J. Cohen, P. Cohen, S. G. West and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Routledge, New York, NY, 3rd ed., 2015.
58. M. Wilson, *Constructing Measures: An Item Response Modeling Approach*, Lawrence Erlbaum, Mahwah, NJ, 2005.
59. J. Cohen, A power primer, *Psychological Bulletin*, **112**(1), pp. 155–159, 1992.
60. P. M. Bentler, Comparative fit indexes in structural models, *Psychological Bulletin*, **107**(2), pp. 238–246, 1990.
61. D. Iacobucci, Structural equations modeling: Fit indices, sample size, and advanced topics, *Journal of Consumer Psychology*, **20**(1), pp. 90–98, 2010.
62. Physport, https://www.physport.org/assessments/FCI. Accessed 09 February 2018.
63. D. G. Johnson, Sorting out the question of feminist technology, in Linda L. Layne, Sharra L. Vostral, and Kate Boyer (eds), *Feminist Technology*, University of Illinios Press, Urbana, IL, pp. 36–54, 2010.
64. J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, *American Journal of Physics*, **78**(10), pp. 1064–1070, 2010.

Appendix A. List of acronyms

Acronym	Definition	Description
aDCI	Abbreviated Dynamics Concept Inventory	Selection of 12 items from the DCI
CFA	Confirmatory factor analysis	Method for testing latent structures
CFI	Comparative fit index	Goodness of fit statistic
CI	Concept inventory	Usually multiple-choice tests that require little or no calculations
DCI	Dynamics Concept Inventory	29-item dynamics concept inventory
df	Degrees of freedom	Measure of how much data is available relative to how many model parameters are being estimated
DIF	Differential item functioning	Scenario of an item functioning differently for distinct groups
FCI	Force Concept Inventory	Physic concept inventory
IRT	Item response theory	Method of modeling latent ability and item characteristics
MG-CFA	Multiple-group confirmatory factor analysis	Method for testing the invariance of a measurement model across multiple groups
RMSEA	Root-mean square error of approximation	Goodness of fit statistic
χ^2	Chi-squared test statistic	Goodness of fit statistic
3PL	Three parameter model	IRT method that models an items difficulty, discrimination, and guessing parameter

Nick A. Stites is a PhD candidate 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.

Kerrie A. Douglas is an Assistant Professor of Engineering Education at Purdue University. Her research is focused on assessment design, methods of validation and fair assessment for diverse engineering learners. This focus includes what evidence and rationale are used to justify assessment use and the consequences of that intended use. She earned her PhD in Educational Psychology, with a concentration on evaluation and assessment, from Purdue University in 2012. She was awarded as a New Faculty Fellow in 2018 at the Frontiers in Education Conference.

David Evenhouse is a graduate student pursuing his PhD in Engineering Education at Purdue University. He graduated from Calvin College in the Spring of 2015 with a BSE concentrating in Mechanical 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.

Jeffrey F. Rhoads is a 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 BS, MS, and PhD 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 thermomechanics of energetic materials; additive manufacturing; and mechanics education. Dr. Rhoads is a Member of the American Society for Engineering Education (ASEE) and a Fellow of the American Society of Mechanical Engineers (ASME), where he serves on the Design Engineering Division's Technical Committee on Vibration and Sound. Dr. Rhoads is a recipient of numerous research and teaching awards, including 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), Robert W. Fox Outstanding Instructor Award, and B.F.S. Schaefer Outstanding Young Faculty Scholar Award; the ASEE Mechanics Division's Ferdinand P. Beer and E. Russell Johnston, Jr. Outstanding New Mechanics Educator Award; and the ASME C. D. Mote Jr., Early Career Award. In 2014 Dr. Rhoads was included in ASEE Prism Magazine's 20 Under 40.

Jennifer DeBoer is an Assistant Professor of Engineering Education and Mechanical Engineering (by courtesy) at Purdue University. Her research focuses on international education systems, individual and social development, technology use and STEM learning, and educational environments for diverse and marginalized learners. She serves as associate editor for the IEEE Transactions on Education. During her first year as assistant professor, she received the NSF's prestigious Early CAREER Award, and in 2017, she received the American Society for Engineering Education Mara Wasburn Early Engineering Educator Award.

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.