Classical Test Theory Analysis of the Dynamics Concept Inventory

Natalie Jorion, Brian Self, Katie James,
Lianne Schroeder, Lou DiBello, Jim Pellegrino

University of Illinois, Chicago/ California Polytechnic State University/
University of Illinois, Chicago/ University of Illinois, Chicago/ University of Illinois, Chicago

Abstract

The Dynamics Concept Inventory (DCI) is an instrument designed to measure students' conceptual understanding of dynamics. Its primary intended use is to examine the effectiveness of teaching practices for helping students overcome misconceptions in the domain, based on evidence of student understanding. Given that many instructors are administrating this assessment in their classrooms, it is important to determine how well the instrument functions relative to the claims of its developers and relative to its intended uses. A further interest is to provide guidance for improving the instrument by identifying aspects of the instrument that may be modified or enhanced. Multiple analyses were conducted for data from two administrations of the instrument using classical test theory. These analyses provide insight into the DCI's conceptual content, measurement properties, and relative validity given its intended use. Overall, evidence shows that the instrument is well suited for low stakes formative assessment use but may have limitations for high stakes uses in its current form. Guidance is provided for the effective implementation and interpretation of the instrument for this purpose. Recommendations are also suggested for future iterations of the instrument and to provide evidence for the resultant changes in measurement properties.

Background

In undergraduate engineering courses, professors often stress procedural problem solving over conceptual understanding of the domain (Miller et al., 2005; Minstrell, Ruth Anderson, & Li, 2011). As a result, engineering students are able to complete courses with high marks while failing to achieve understanding of key concepts and simultaneously maintaining problematic misconceptions in the domain. Some of these misconceptions can be attributed to faulty preconceptions and inflexible knowledge transfer stemming from instructional examples (Ruiz-Primo et. al, 2012). Concept Inventories (CIs) have been cited as a means to address this instructional dilemma. CIs are low-stakes, multiple-choice assessments that purport to measure students' conceptual understanding in a discipline. The multiple-choice distractors are based on common student mistakes, which are often derived by developers from student responses to open-ended questions. Many instructors are using these instruments to make inferences regarding students' understanding of the domain, with the goal of improving future teaching.

One such CI is the Dynamics Concept Inventory (DCI), which claims to assess concepts that instructors perceive as difficult for students (Gray et. al, 2005). This test has promise as a tool to help improve instruction by providing evidence of student thinking. An important step in supporting instructional decisions with student data is validating the test's measurement properties. In this paper, Classical Test Theory (CTT) is used to investigate functioning of the questions of the inventory and of the inventory as a whole. These analyses can help instructors interpret students' scores on the DCI. In addition, we substantiate various aspects of the validity of the DCI for particular kinds of classroom use and provide some ideas of how the inventory may be modified or improved.

Several analyses have been conducted on previous versions of the DCI for the purpose of selecting items for the current instrument. These analyses include basic reliability tests (Cronbach's alpha) and distractor analyses (Gray et al., 2005). The current study differs from earlier research in several regards. First, it analyzes performance on the most current version of the DCI. Second, it examines how the instrument functions overall and explains how these findings affect interpretation and use of inventory outcomes. In addition to reliability analyses, a measure of standard error is presented with an explanation of how that informs interpretation of total scores. Third, the study examines how the items are functioning in the context of the instrument as a whole, which can help users understand the utility of the items for informing and refining instruction.

Method

Participants

The analyses made use of post-test DCI data from students at two large public universities. One of these schools is on the semester system, while the other is on the quarter system. The students took the test for an undergraduate dynamics course during June, 2011. The majority of these students were sophomore engineering majors, including mechanical, civil, aero, biomedical, and industrial engineering. The combined datasets totaled 966 cases.

Instrument

The version 1.0 of the DCI that was analyzed has 29 questions, five of which are taken without change from the Force Concept Inventory (FCI). The developers provided a list of 14 conceptual categories for the inventory questions, ranging from one to five items per category (see Appendix for a list of the concepts and assignment of items).

Procedures

We used CTT to investigate the extent to which the overall test total score is reliable and whether there are particular items that seem to function differently than the rest of the test. CTT assumes that for a given assessment each examinee possesses a "true score" and that each observed score is measured as the true score plus error:

$$X = T + \varepsilon$$

where *X* represents the observed score, T represents the true score, and ε represents the error. The true score can be understood conceptually as the examinee's average observed scores on the same assessment over an infinite number of times (assuming no test-retest effects).

Since the true score cannot be observed directly, various approaches have been developed to estimate the reliability of the observed total score as a relationship between the observed score and the true score. Cronbach's alpha, in particular, measures internal consistency of the individual item scores that make up the total score. Alpha can range from 0 to 1; an alpha close to 1 indicates that the items are closely related as a group, suggesting a dominant underlying construct. The formula for Cronbach's alpha is defined as follows:

$$\alpha = \frac{N\overline{c}}{(\overline{v} + (N-1)\overline{c})}$$

where *N* is the number of items, $\bar{c}$ is the average inter-item covariances among all item pairs, and $\bar{v}$ is the average of the item variances. The formula for alpha is sensitive to the number of items in the assessment instrument. Typically the greater the number of items on a test, the greater the value for alpha. Values of alpha greater than 0.7 are acceptable, and measures between 0.8 and 0.9 are desirable (Nunnally& Bernstein, 1994).

The standard error of estimation enables us to define a confidence interval for a student's true score given her observed scores. The measure uses the standard deviation of the test scores and the overall test reliability:

$$SEE = (S_X)(\sqrt{r_X})\left(\sqrt{1 - r_X}\right)$$

where r_X is the reliability coefficient and S_X is the standard deviation of test scores (Harvill, 1991). The greater the test reliability, the smaller the standard error of measurement is. A test with perfect reliability—that is, a Cronbach's alpha of 1—would have a standard error of estimation of 0 and a true score equal to the observed score. Using this formula, we can derive the approximate 68% confidence interval for a student's true score around an observed total score. For example suppose an examinee's observed score is X. Then the 68% confidence interval for that student's true score is:

$$CI = \left(\overline{X} + (r_X)\left(X - \overline{X}\right)\right) \pm SEE$$

where $\bar{X}$ is the mean score for a reference group and *X* is the obtained test score, and SEE is the standard error of estimate as given above. This formula indicates the confidence intervals for the unobserved true score around the observed score X. In other words, for a given Cronbach's alpha and related standard error of estimate as computed above, this formula provides a confidence interval for the student's true score, given their observed score. A specific example of application of this statistic is presented below in the Results.

Functioning of individual items was evaluated using three measures. First, a quantity called "*Cronbach's alpha if-item-deleted*" was used which is calculated exactly the same way as the Cronbach alpha for total score, except it is calculated on the set of all items except for the item being deleted. This measure is compared with the alpha of the overall test. Since this is a Cronbach's alpha, it ranges from 0.0 to 1.0. As previously mentioned, the number of items on a

test directly affects Cronbach's alpha; the more items on a test, the greater the value of alpha should be, assuming uniform quality of the items. Deleting an item from an assessment should, in theory, result in a lower alpha. Thus, items that have a *higher* alpha if-item-deleted score than the overall alpha they are likely detracting from the assessment's overall reliability. This may be because an item is measuring a different construct from the rest of the test or simply because the item is a poor measure of the ability measured by the remainder of the items in the test.

The second statistic investigated was *item difficulty*. A given question's difficulty (or p-value) is the proportion of examinees that answered the question correctly. The item difficulty value also ranges from 0.0 to 1.0. The greater the value, the easier the item is. Ideally, this statistic should be between 0.2 and 0.8 for an item to be sufficiently informative. There are several possible explanations for items that have a measure of less than 0.2: the item may be too difficult relative to the sample tested, the item may not be worded clearly, or there may be more than one correct answer. We further investigated particular items that fell out of this range. We also created distributions of answer choices selected for each item, noting any instances where there a distractor was chosen more frequently than a correct item choice.

The third measure used was *item discrimination*, which is calculated as the point-biserial correlation between item score and total score. This is just the ordinary Pearson correlation between a 0/1 item score and the total score excluding the item of interest. This statistic indicates the extent to which an item discriminates between students with higher and lower total scores. In other words, an item with a greater discrimination is more frequently answered correctly by students with a "high" level of knowledge than by those with a "low" level of knowledge. This statistic can range from -1.0 to 1.0, with negative values indicating items that were more frequently answered correctly by students with a lower ability. An item's discrimination should be greater than 0.2. Items with a value lower than 0.2 may be testing a different construct than the rest of the test, may have a seductive distractor selected more frequently by those students with a higher ability, or may be a poor item for other reasons.

Results

The mean of the total scores on the DCI for all students from the given sample was 14.27 (*SD*=4.59) with a maximum possible score of 29. Figure 1 shows the distribution of student scores. The overall Cronbach's alpha for this test is 0.744, which is a modest reliability measure for a test with 29 items intended to be used for formative assessment or for instructional evaluation purposes. Given the standard deviation and the reliability, the standard error of estimation for the sample is 2.00. We can use the formula given above to illustrate the effect of standard error of estimation for interpreting student total scores. For example, suppose a given student has total score close to the mean score, for example total score 14.

Applying the formula given above, the 68% confidence interval is defined to be:

$$\left(\overline{X} + (r_X)\left(X - \overline{X}\right)\right) \pm SEE =$$
$$\left(14.52 + (0.744)(14 - 14.52)\right) \pm 2.00 =$$
the interval from 12.07 to 16.07.

Thus there is a 68% chance that the student's true score is between 12 and 16 (Harvill, 1991). Confidence bands that do not overlap have a high probability of being distinct from each other. By contrast, two students with scores 12 and 16 cannot be inferred with 68% confidence to have different true scores. With the same mean and standard deviation of total scores, for the standard error of estimation to decrease to 1, the reliability measure would have to increase to 0.95. However, it should be noted that the effective standard error of estimate for the observed score increases the further the observed score is from the mean. More sophisticated measures exist for calculating the standard error of estimation which use item response theory and other model-based approaches (Baker, 2001; Crocker & Algina, 2006; Hambleton et al., 1991).

Individual item analyses are presented in Table 1. Bolded items denote measures that are outside of the recommended range for each statistic. The majority of the items had adequate item discrimination measures over 0.2. Items with difficulty measures of less that 0.2 (i.e., that were difficult for this sample of examinees) are Q3, Q5, and Q29. Items that had difficulty measures of more than 0.8 (i.e., that were somewhat easy for this group) are Q1, Q7, and Q14. Several questions have higher alphas if-item-deleted scores: Q5, Q10, Q13, and Q23, suggesting that they are internally inconsistent with the other items of test and might not cohere well conceptually with the rest of the test. The item discriminations and difficulties are fairly well distributed, indicating that the questions are suitable to assess a wide range of conceptual mastery of the domain (see Figure 2).

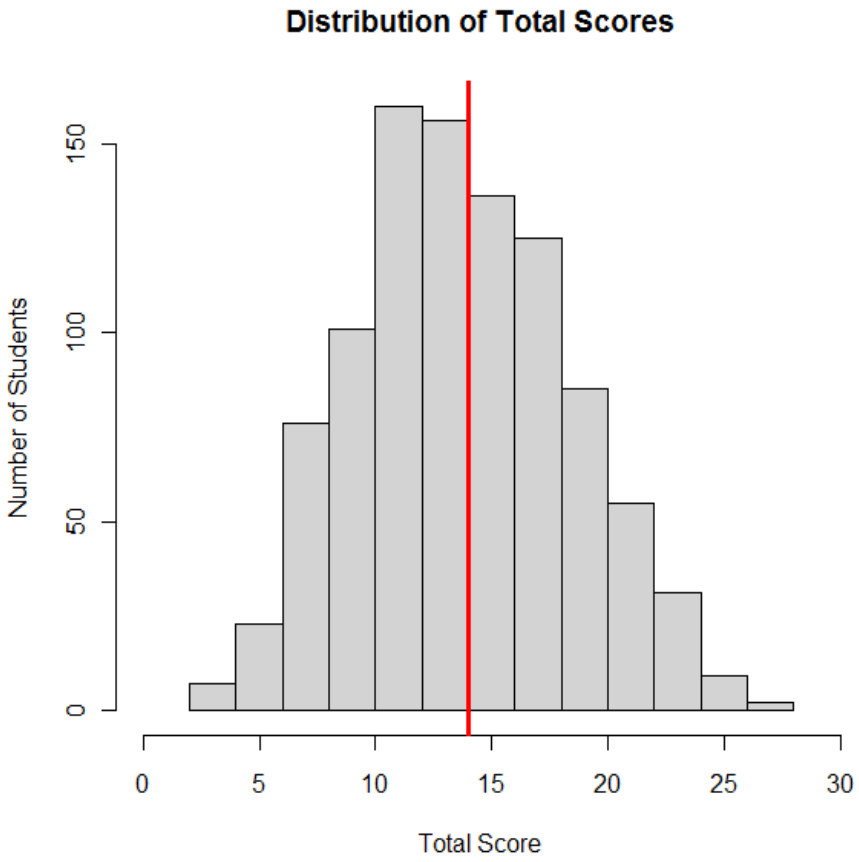

Figure 1. *Total Score distribution (n=966). Red line indicates median score.*

Table 1. *Item statistics for the DCI.*
Measures in bold indicate either a discrimination measure below 0.2, a difficulty measure outside the range of 0.2-0.8, or a Cronbach's alpha-if-item-deleted of more than the overall test reliability (0.744).

Item	Discrimination	Difficulty	Cronbach's Alpha If Item Deleted
Q1	**0.196**	**0.891**	0.740
Q2	**0.145**	0.639	0.744
Q3	0.255	**0.133**	0.738
Q4	0.403	0.609	0.728
Q5	**-0.004**	**0.058**	**0.747**
Q6	0.338	0.525	0.732
Q7	0.262	**0.839**	0.737
Q8	0.370	0.637	0.730
Q9	0.341	0.371	0.732
Q10	**0.080**	0.443	**0.749**
Q11	**0.192**	0.332	0.741
Q12	0.223	0.421	0.739
Q13	**0.104**	0.355	**0.747**
Q14	0.248	**0.913**	0.738
Q15	0.326	0.642	0.733
Q16	0.385	0.729	0.729
Q17	0.403	0.715	0.728
Q18	0.248	0.429	0.738
Q19	**0.183**	0.273	0.741
Q20	0.215	0.684	0.740
Q21	0.363	0.322	0.730
Q22	0.395	0.588	0.728
Q23	**0.114**	0.360	**0.746**
Q24	0.360	0.536	0.730
Q25	0.286	0.520	0.735
Q26	0.200	0.294	0.740
Q27	0.358	0.431	0.731
Q28	0.222	0.402	0.739
Q29	0.262	**0.177**	0.737

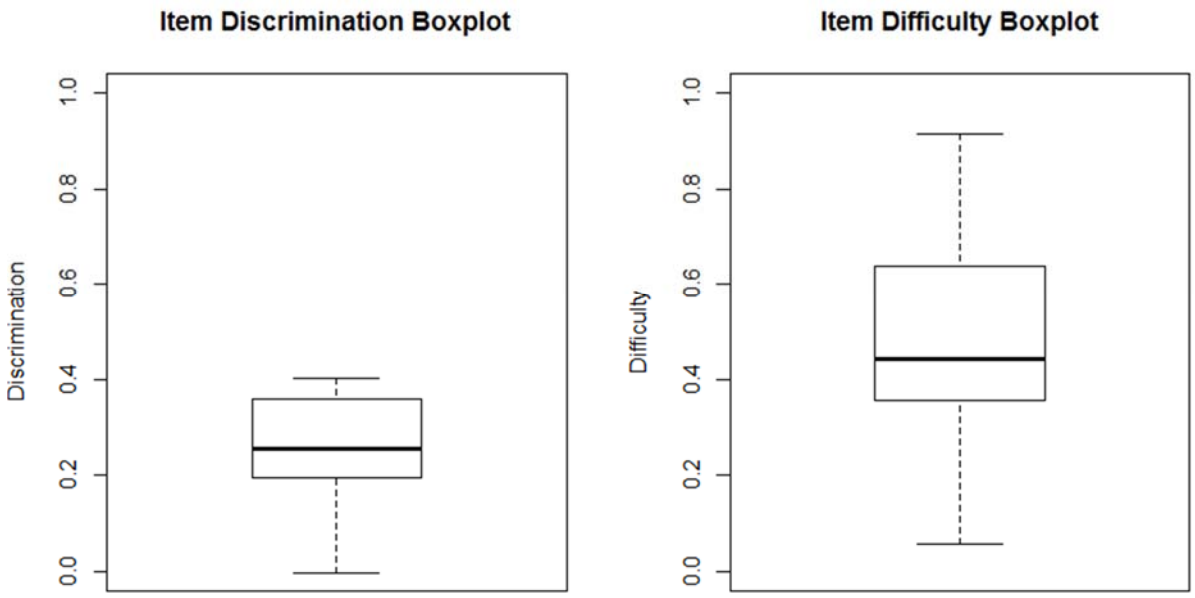

Figure 2. *Range of item discriminations and difficulties. Note that discrimination values can be negative, with minimum possible value of -1; 0 is the minimum here for comparative purposes.*

Discussion

The analyses indicate that as a whole, the DCI is a fairly reliable instrument with a reasonable standard error of estimation in the score distribution close to the mean. Even though the test purposefully covers a wide range of concepts within the domain of dynamics, the reliability measure suggests a relatively cohesive assessment. The standard error indicates the degree of confidence instructors can have in students' observed scores and differences between bands of observed scores.

It should be noted that not all the items are contributing equally to its alpha measure. Those items with values outside the desired range of difficulty and those with low inter-item covariance do not add as much to the overall reliability. However, this by itself does not negate the value of items that fall outside this range. Easy items could be retained if it is felt that they can indicate whether students have mastered the most fundamental concepts. That argument may be made in particular for the items drawn from the FCI (Q1, Q7, Q14, Q15, and Q16). Items with difficulties of less than 0.2 may be included in an inventory for their value in indicating understanding of more challenging concepts. In this way, both the interpretation of the DCI in its existing form as well as the decision of which items are candidates for enhancement or replacement can be informed by the variety of analyses reported here as well as possible instructional considerations give the individual item data.

The results suggest how engineering instructors could appropriately leverage results from the DCI to inform instruction. The DCI would be well-suited for assessment at the student-level and the classroom-level. On an individual-level, the test could indicate which students still harbor misconceptions about dynamics. Instructors could then create interventions as appropriate. On a classroom-level, the DCI could indicate whether instruction of particular concepts has been

effective. If instructors find that certain misconceptions are prevalent among the majority of their students on a post-test, they may want to investigate the corresponding lessons. In some cases, instruction may help to dispel a problematic preconception, but then incite another misconception.

Although classical test theory can provide informative measures of test functioning, it does have limitations. For example, CTT measures are less accurate for student scores at the extreme ends of the total distribution. To further understand the performance of the inventory and the interpretation of test and item performance we are currently using different measurement models such as Item Response Theory (IRT). We are also using IRT, factor analysis, and structural equation modeling to uncover the DCI's underlying structure. Such analyses can be helpful for determining whether the sub-scores of the inventory have meaningful interpretations.

Conclusion

In this study, Classical Test Theory was used to investigate the measurement properties of the most current version of the DCI. To analyze the functioning of the DCI as whole, the focus was on reliability measures and the standard error of estimation. These measures can indicate the degree of confidence one can have regarding the relationship between students' observed scores and their true scores. Individual item analyses were used to show how they were functioning within the larger assessment. The wide range of item difficulty can help instructors differentiate between varying degrees of dynamics conceptual mastery. Overall, the results indicate that the DCI can be a valuable low-stakes instrument that professors can use to identify conceptual mastery of dynamics. In addition, the specific results suggest areas in which improved functioning of the instrument may be possible by enhancing or replacing some items.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation under Grant Nos. 0918552, 0920589, 0815065. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Baker, F. (2001). *The basics of item response theory.* ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD. Retrieved October 30, 2004, from <http://echo.edres.org:8080/irt/baker/http://edres.org/irt/baker/>

Crocker, L. & Algina, J. (2006). *Introduction to classical and modern test theory*. New York: Wadsworth Publishing Co.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Measurement Methods for the Social Science).* Newbury Park, CA: Sage Publications.

Gray, G., Evans, D., Cornwell, P., Costanzo, F., & Self, B. (2005). *The Dynamics Concept Inventory Assessment Test: A Progress Report*, Proceedings of the 2005 American Society for Engineering Education Annual Conference, Portland, OR.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 33-41.

Miller, R. L., Streveler, R., Olds, B., & Nelson, M. (2005). Concept Inventories Meet Cognitive Psychology: Using Beta Testing as a Mechanism for Identifying Engineering Student Misconceptions. Proceedings of the American Society for Engineering Education Annual Conference (electronic), Portland, Oregon.

Minstrell, J., Anderson, R., & Li, M. (2011, May). Building on Learner Thinking: A Framework for Assessment in Instruction. workshop of the committee on Highly Successful Schools or Programs for K-12 STEM Education. National Research Council, Washington, DC, May.

Nunnally, J. C. & Bernstein, L. (1994). *Psychometric Theory* (3rd ed.) McGraw-Hill, New York.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691-712.

Appendix: Grouping of DCI Concepts

Concepts (from 2005 ASEE paper)

Concept	Question
1. Different points on a rigid body have different velocities and accelerations, which vary continuously.	2,3
2. 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.	7,11,15,16, 24
3. Angular velocities and angular accelerations are properties of the body as a whole and can vary with time.	4, 6
4. Rigid bodies have both translational and rotational kinetic energy.	10
5. The angular momentum of a rigid body involves translational and rotational components and requires using some point as a reference.	25, 26
6. Points on an object that is rolling without slip have velocities and acceleration that depend on the rolling without slip condition.	21, 22, 23
7. In general, the total mechanical energy is not conserved during an impact.	18, 20
8. An object can have (*a*) nonzero acceleration and zero velocity or (*b*) nonzero velocity and no acceleration.	9, 23
9. The inertia of a body affects its acceleration.	12,13,17
10. The direction of the friction force on a rolling rigid body is not related in a fixed way to the direction of rolling.	27,28
11. A particle has acceleration when it is moving with a relative velocity on a rotating object.	5, 19
12. An object moving in a curved path always has a normal component of acceleration	8
13. The direction of the friction force between two objects depends on their relative velocity or their tendency for relative motion.	29
14. Newton's third law dictates that the interaction forces between two objects must be equal and opposite.	1,14