ANALYZING THE CALCULUS CONCEPT INVENTORY: CONTENT VALIDITY, INTERNAL STRUCTURE VALIDITY, AND RELIABILITY ANALYSIS

Jim Gleason
The University of Alabama
jgleason@ua.edu

Matt Thomas
Ithaca College
mthomas7@ithaca.edu

Spencer Bagley
University of Northern Colorado
Spencer.Bagley@unco.edu

Lisa Rice
Arkansas State University
lrice@astate.edu

Diana White
University of Colorado Denver
diana.white@ucdenver.edu

Nathan Clements
University of Wyoming
nathan.clements@uwyo.edu

We present findings from an analysis of the Calculus Concept Inventory. Analysis of data from over 1500 students across four institutions indicates that there are deficiencies in the instrument. The analysis showed the data is consistent with a unidimensional model and does not have strong enough reliability for its intended use. This finding emphasizes the need for creating and validating a criterion-referenced concept inventory on differential calculus. We conclude with ideas for such an instrument and its uses.

Keywords: Post-Secondary Education; Assessment and Evaluation; Research Methods; Instructional Activities and Practices

Introduction and Literature Review

As educators and educational researchers, we seek to develop calculus courses effective in building conceptual understanding in addition to procedural fluency, and continually investigate promising new pedagogical strategies. The Mathematical Association of America recommends that all math courses should build conceptual understanding, "mental connections among mathematical facts, procedures, and ideas" (Hiebert & Grouws, 2007, p. 380) by helping "all students progress in developing analytical, critical reasoning, problem-solving, and communication skills and acquiring mathematical habits of mind" (Barker et al., 2004, p. 13).

Concept inventories have emerged over the past two decades as one way to measure conceptual understanding in STEM education. These inventories intend to assess student understanding of concepts before entering a course addressing those concepts. Thus, students are required to use common sense and prior knowledge to respond to assessment items. After completing a course, the concept inventory can measure gains in conceptual understanding. Therefore, items should avoid using terminology taught in the course to which students have no prior exposure.

The first concept inventory to make a significant impact in the undergraduate education community was the Force Concept Inventory (FCI), written by Hestenes, Wells, & Swackhamer (1992). Despite the fact that most physics professors considered the Inventory questions "too trivial to be informative" (Hestenes et al., 1992, p. 2) at first glance, students did poorly on the test, and comparisons of high-school students with university students showed modest gains between the two. Of the 1,500 high-school students and over 500 university students who took the test, high school students were learning 20%-23% of the previously unknown concepts, and college students at most 32% (Hestenes et al., 1992, p. 6). Through a well-documented process of development and refinement, the test has become an accepted and widely used tool in the physics community, and has led to changes in the methods of instruction for introductory physics.

The FCI paved the way for the broad application of analyzing student conceptual understanding of the basic ideas in a STEM subject area (Hake, 1998, 2007; Hestenes et al., 1992). Concept inventories exist in a variety of scientific disciplines; including physics, chemistry, astronomy, biology, and geoscience (Libarkin, 2008).

Bartell, T. G., Bieda, K. N., Putnam, R. T., Bradfield, K., & Dominguez, H. (Eds.). (2015). *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. East Lansing, MI: Michigan State University

More recently, Epstein (2007, 2013) developed the Calculus Concept Inventory (CCI) for introductory calculus. However, there is a lack of peer-reviewed literature on its development or psychometric analysis. Additionally, several recent analyses call into question the ability of the CCI to measure conceptual understanding. One study showed that the current CCI measured no difference in conceptual understanding between students in a conceptually focused class with frequent student group work and those in a traditional lecture based class, even though other measures indicated that a difference existed (Bagley, 2014). While this result may reflect shortcomings of the conceptually focused class, it may also suggest the inadequacy of the CCI.

The concerns with the CCI motivated us to take a deeper look at how it performs for its original purpose. Specifically, we wanted to determine if the CCI measured gains in conceptual knowledge and to investigate its reliability. In this study, we analyze the results on the CCI from over 1500 students at four institutions to determine whether there is evidence that the CCI, in its current form, exhibits the psychometric properties originally suggested by the author, and to suggest appropriate potential modifications or revisions.

Calculus Concept Inventory

Content Validity

The Calculus Concept Inventory (CCI) was developed by a group of seven individuals to measure topics from differential calculus that they believed were basic constructs (Epstein, 2013). The main purpose of the instrument is to measure classroom normalized gains (change in the class average divided by the possible change in the class average) for the purpose of evaluating the impact of teaching techniques on conceptual learning. The developers of the instrument intended the instrument to measure above random chance at the pre-test setting and to avoid "confusing wording" (Epstein, 2013, p. 7). However, a released CCI item uses terminology, including "derivative" and "f'(x)" (Epstein, n.d.), which is not part of the vocabulary of a first-time calculus student. Such items would be confusing to the student and generate responses around random chance for those items. We seek to determine the extent of the use of such terminology to verify that vocabulary issues do not confound results from the CCI.

Internal Structure Validity

The dimensionality of the CCI is unknown. Epstein (2013) states that the instrument has two primary components, related to functions and derivatives, with a third dimension related to limits, ratios, and the continuum. However, the use of a total percent correct to determine normalized gains implies that the instrument measures a single construct evenly distributed over the 22 items. These two proposed structures of the instrument are contradictory and no details regarding the analysis conducted to support the three-component structure exist. A comprehensive analysis of the internal structure of the instrument is thus necessary to determine whether a unidimensional model is appropriate.

Reliability

Epstein (2013) reports that the CCI has an internal consistency reliability 0.7 for Cronbach's alpha. This level of internal consistency is at the low end of an acceptable range for an instrument designed to measure differences in means between groups of at least 25-50 individuals. However, there is no such standard for internal consistency necessary for comparing the normalized gains of two different groups. In fact, the use of the normalized gain as a measurement parameter is questionable (Wallace & Bailey, 2010). Instead, the similar types of gains can be measured using ability estimates obtained through item response theory models. Therefore, there is a need to use such models to determine the internal consistency reliability of the CCI.

Bartell, T. G., Bieda, K. N., Putnam, R. T., Bradfield, K., & Dominguez, H. (Eds.). (2015). *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. East Lansing, MI: Michigan State University

Methods

Content Validity

Since the CCI was designed to measure normalized gains in conceptual knowledge of calculus, it is given as both a pre-test and a post-test. As such, at both of the sittings, the test should measure conceptual understanding, and not include items requiring vocabulary and notation specific to calculus. Otherwise, students who are repeating calculus would likely have higher pretest scores, regardless of their conceptual understanding of calculus, and would thus likely have lower normalized gains, as seen in previous studies (Epstein, 2013). Therefore, we conducted an analysis of the items to determine which items may contain vocabulary and/or notation not included in any of the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010) that have become accepted as preparation for calculus throughout most of the United States.

Internal Structure Validity

We collected data from approximately 2000 students at four universities at the beginning and the end of a first semester calculus class. We cleaned the data by eliminating subjects with missing data and randomly selecting either a pre-test or post-test for all remaining subjects to avoid dependent samples. This left a sample size of 1792 students with an even distribution of pre-tests and post-tests.

We then used the eigenvalues of the inter-item correlation matrix to determine the expected number of factors related to the instrument, followed by a confirmatory factor analysis based on the predicted number of factors, with a bent toward a unidimensional model. In the eigenvalue analysis, we compared the results from the actual data to results from randomly generated data with the same sample size and with a 20% probability of correct answers, as nearly all of the items on the CCI had five choices.

Reliability

Using the results of the factor analysis, we used an appropriate unidimensional or multidimensional item response theory model to analyze the internal reliability of the instrument and to measure the test information and standard errors for the instrument.

Results

Content Validity

Out of the 22 items on the CCI, nine contained language or notation not included in any standards for courses that are considered prerequisite for calculus. These included the words derivative and concavity, and notation such as $f'(x)$, $f''(x)$, and dy/dx. An additional two items contained language closely related to some precalculus topics; for instance, some students may have exposure to the relationship between velocity and acceleration and the concept of linear approximations. However, these topics are not necessarily included in the courses prior to calculus.

Therefore, the CCI does not satisfy the conditions necessary to measure conceptual understanding for students as they enter a calculus course. However, since all of these language and notation conventions are part of the normal language during the first semester of calculus, including such language on a test at the end of a semester of calculus may measure conceptual understanding. This issue needs to justification by anyone using the standard normalized gains when researching or evaluating first semester calculus courses.

Internal Structure Validity

From the analysis of the eigenvalues from the factor analysis, the CCI has at most two components. Both the first and the second eigenvalue are above the 95% confidence interval for the randomly generated data. However, since the second eigenvalue (1.24) is extremely close to the 95%

Bartell, T. G., Bieda, K. N., Putnam, R. T., Bradfield, K., & Dominguez, H. (Eds.). (2015). *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. East Lansing, MI: Michigan State University

confidence interval of the eigenvalue generated by random data 1.1765 +/- 0.04, this second component may or may not actually be present (since a large first eigenvalue will pull up the second eigenvalue).

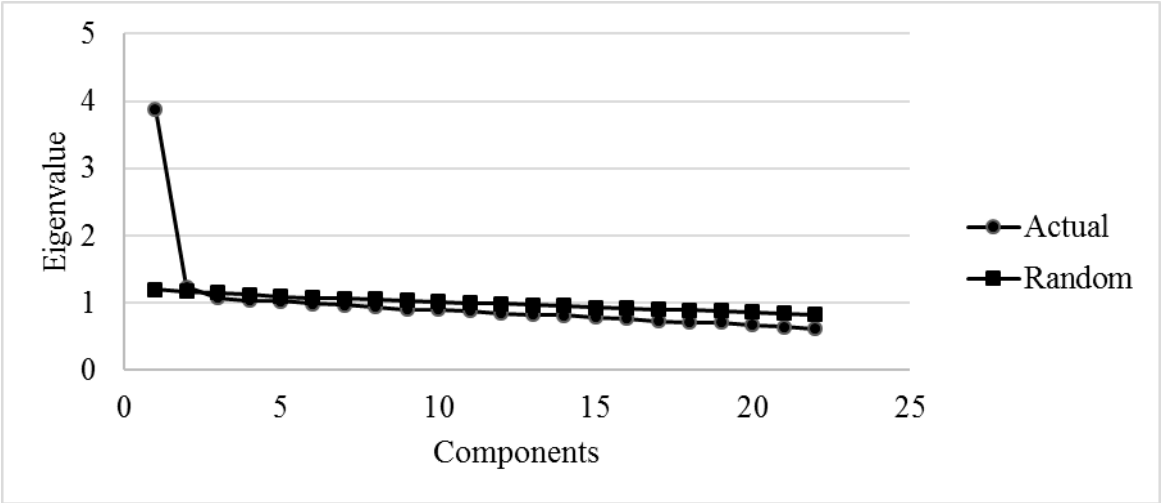

Figure 1: Scree Plot for Calculus Concept Inventory

Table 1: Item CFA Estimates for CCI

	Full CCI		Abbreviated CCI	
Item	Estimate	Standard Error	Estimate	Standard Error
Question 1	1.000			
Question 2	5.776	1.313	1.000	
Question 3	5.537	1.264	0.961	0.065
Question 4	5.649	1.288	0.978	0.065
Question 5	4.574	1.058	0.802	0.062
Question 6	3.243	0.769	0.560	0.053
Question 7	3.497	0.825	0.604	0.055
Question 8	5.055	1.158	0.877	0.062
Question 9	4.792	1.103	0.830	0.062
Question 10	4.693	1.084	0.803	0.062
Question 11	0.816	0.349		

Question 12	2.380	0.610	0.414	0.056
Question 13	4.228	0.985	0.735	0.060
Question 14	3.386	0.803	0.587	0.055
Question 15	3.735	0.880	0.650	0.058
Question 16	2.928	0.704	0.504	0.052
Question 17	5.619	1.282	0.975	0.065
Question 18	1.535	0.412		
Question 19	3.322	0.790	0.570	0.055
Question 20	3.575	0.849	0.617	0.058
Question 21	3.857	0.917	0.661	0.064
Question 22	3.885	0.913	0.678	0.059

Since the scree plot and eigenvalue analysis favors a unidimensional structure, and since the intended use of the instrument is as a one-dimensional inventory, a one-dimensional confirmatory factor analysis model was used to determine model-data fit. The model had 231 degrees of freedom, $p<0.001$, with the item estimates given below. The fit indices were excellent with a Comparative Fit Index (CFI) of 0.936 and a Root Mean Square Error of Approximation (RMSEA) of 0.024 (Hu & Bentler, 1999). Therefore, a unidimensional model is assumed to fit the data well. One notices that three of the items (1, 11, and 18) have significantly lower estimates that the remaining items. If one removes these items, we maintain the unidimensionality of the instrument (CFI: 0.939 and RMSEA: 0.028) and all estimates are approximately equal values. This enables a more appropriate use of number correct to estimate an individual's ability without having to scale the values of certain items.

Reliability

Since the instrument satisfies the unidimensionality assumption, one, two, or three-parameter models can be used to analyze the data. Since the different items are believed to have different discrimination, only the two and three parameter models were used. The three-parameter model did not have good model-data fit on several of the items loading heavily on the construct with the c parameters for the majority of the items significantly below random chance. Therefore, a two-parameter model was determined to be the best fit of the data and the theoretical construct of the inventory. In the analysis of the two-parameter model, three items demonstrated a weak fit, items 1, 11, and 18. These three items also had low loadings in the factor analysis and so were removed from the analysis to determine if the remaining items have an improved fit. The remaining 19 items had a good fit (-2LL of 37258, $p<0.0001$) with the two parameter model. The standard error for the ability estimate of individuals is extremely high with the lowest value of 0.4128 logits and an average error of 0.7307 logits (see Figure 2). For example, if an individual is at the mean in terms of actual conceptual understanding of calculus, as measured by the CCI, the measured score of the person by the inventory has a 68% chance of being within 0.42 logits of the mean. Therefore, the inventory would only be able to differentiate between samples of means if there is a substantial difference

between the samples or the sample size approaches 100 students each. Furthermore, in order to use the logit scores one must first transform the percent correct score into logit scores using the results in Table 2.

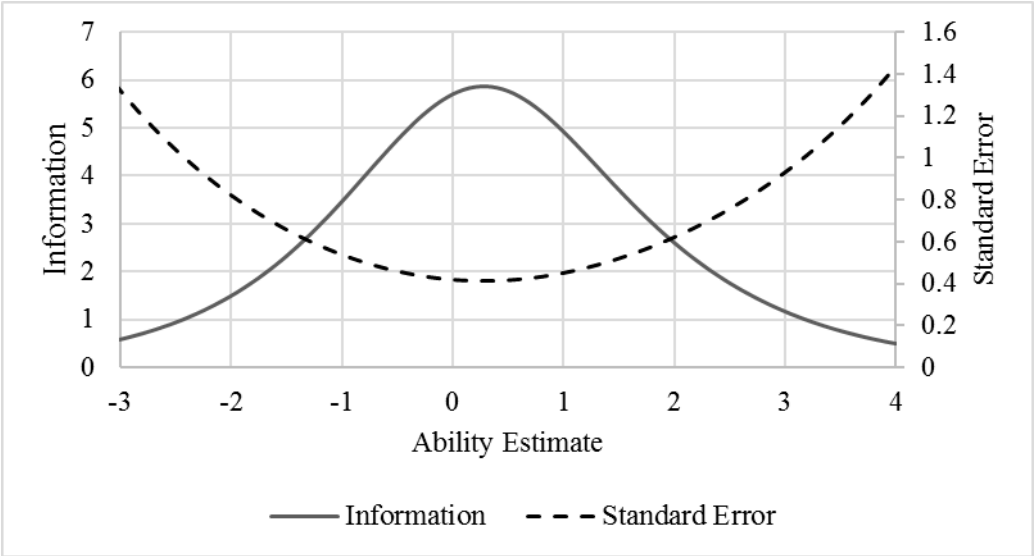

Figure 2: Test Information Function and Standard Error

Table 2: Transformation of Scores to Logits

Number Correct	Ability Estimate		Number Correct	Ability Estimate		Number Correct	Ability Estimate
0	-3.05		7	-0.13		14	1.47
1	-2.52		8	0.10		15	1.77
2	-1.75		9	0.31		16	2.13
3	-1.28		10	0.53		17	2.60
4	-0.93		11	0.74		18	3.36
5	-0.63		12	0.97		19	4.00
6	-0.37		13	1.21			

Discussion and Conclusion

The purpose of this study was to assess the degree to which the CCI conforms to certain standards for psychometric properties, including content validity, internal structure validity, and reliability. We conclude that the existing CCI does not conform to accepted standards for educational testing (American Educational Research Association, 2014; DeVellis, 2012). We thus argue that there is a need to create and validate a criterion-referenced concept inventory on differential calculus. Such a concept inventory would significantly impact teaching and learning during the first two years of undergraduate STEM students by providing a resource to measure students' conceptual understanding of differential calculus. The work of Carlson, Madison, and West (2010) in developing

the Calculus Concept Readiness instrument could serve as a model and foundation for a differential calculus concept inventory. Such an instrument would be useful for instructors for formative and summative assessment during their calculus courses to improve student learning. Researchers and evaluators to measure growth of student conceptual understanding could also use such an instrument during a first semester calculus course to compare gains of students in classrooms implementing differing instructional techniques.

Acknowledgements

We are thankful to Guada Lozano and Chris Rasmussen for contributions that have helped us carry out this work.

References

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bagley, S. (2014). *Improving student success in calculus: A comparison of four college calculus classes*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Barker, W., Bressoud, D., Epp, S., Ganter, S., Haver, B., & Pollatsek, H. (2004). *Undergraduate programs and courses in the mathematical sciences: CUPM curriculum guide*. Washington, DC: Mathematical Association of America.

Carlson, M., Madison, B., & West, R. (2010). *The Calculus Concept Readiness (CCR) instrument: Assessing student readiness for calculus. arXiv preprint arXiv:1010.2719*. Retrieved from http://arxiv.org/abs/1010.2719

DeVellis, R.F. (2012). *Scale Development: Theory and applications* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Epstein, J. (2007). Development and validation of the Calculus Concept Inventory. In *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* (pp. 165–170).

Epstein, J. (2013). The calculus concept inventory - Measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society, 60*(8), 2-10.

Epstein, J. (n.d.). Calculus concept inventory instrument. *Field-tested Learning Assessment Guide for science, math, engineering, and technology instructors: Tools*. Retrieved from http://www.flaguide.org/tools/diagnostic/calculus_concept_inventory.php

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74.

Hake, R. R. (2007). Six lessons from the physics education reform effort. *Latin American Journal of Physics Education*, *1*(1), 24–31.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. doi:10.1119/1.2343497

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371-404). Reston, VA: National Council of Teachers of Mathematics.

Libarkin, J. (2008). *Concept inventories in higher education science*. National Research Council Promising Practices in Undergraduate STEM Education Workshop 2, Washington, DC, 13-14 October 2008.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

Wallace, C.S. & Bailey, J.M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review, 9*(1), 010116. doi:10.3847/AER2010024