Moving Beyond Total Scores With Skill-Proficiency Profiles on Concept Inventories

Kevin Roberge
University of Maine

Jayson M. Nissen
Montana State University

Vy Le
Iowa State University

Ben Van Dusen
Iowa State University

Research-based assessments (RBAs) are widely used to evaluate student learning in STEM, yet most provide only total scores. We illustrate the value of moving beyond overall performance by applying a cognitive diagnostic (CD) model to student responses on the Precalculus Concept Assessment (PCA) and the Calculus Concept Inventory (CCI). Using 1,366 pre/post responses collected through the LASSO platform, we modeled five key calculus skills: prerequisites, limits, derivatives, applications of derivatives, and introductory integration. Traditional analyses revealed small but statistically significant score increases (d = 0.09 for PCA; d = 0.26 for CCI). CD modeling also showed limited shifts in skill proficiency profiles: students who lacked prerequisite skills did not gain those skills. These results underscore how CDs provide finer-grained evidence than total scores and motivate the development of a Calculus Cognitive Diagnostic to support timely instructional interventions.

Keywords: Calculus, Assessment, Learning, Cognitive Diagnostic

Research-Based Assessments (RBAs) have become foundational in DBER because they are developed through iterative qualitative and quantitative studies that build a cumulative validity argument aligned to field standards. This systematic approach positions RBAs to measure learning and to inform instructional decision-making at scale. In physics, RBAs enabled early demonstrations that interactive engagement markedly outperforms lecture (e.g., Hake, 1998), guided the design and evaluation of Modeling Instruction (Brewe, 2008), and later underpinned cross-disciplinary meta-analytic evidence that active learning improves achievement and lowers failure rates (Freeman et al, 2014). More recent work shows that although students *feel* they learn less during active engagement, they in fact learn more (Deslauriers et al., 2019), and that lecturing still predominates in undergraduate STEM (Stain et al., 2018). These results underscore the need for sustained, RBA-informed reform, and RBAs designed to meet the needs of educators and researchers pursuing those reform efforts.

Mathematics education researchers have developed at least six RBAs covering precalculus and calculus content (Epstein, 2006: Carlson et al., 2010; Carlson et al., 2015; Thompson et al., 2018; Lee et al., 2025). Researchers tend to use these fixed-length instruments to measure learning across a semester though they have played a smaller role in the RUME community than similar instruments have in physics (e.g. Hake, 1998). This limited use may result from subsequent studies that questioned the psychometric validity arguments of some of these instruments (Gleason et al., 2019; Jones, 2021; Lee et al., 2025).

We are developing a calculus cognitive diagnostic (CCD). Whereas current fixed length RBAs primarily assess overall ability, a cognitive diagnostic (CD) assessment measures student proficiency on individual skills and overall ability (Leighton & Gierl, 2007). The CD classifies students into skill profiles to provide instructors, researchers and students with finer grained information about student learning (de la Torre & Minchen, 2014). Unlike many fixed-length tests, CDs can assess a subset of skills and overall ability whenever needed, even multiple times a semester, without compromising the instrument. By drawing from a larger item bank, CDs also prevent over exposure of items. Our CCD development uses a computerized adaptive testing platform (LASSO, 2025) that optimizes the information each question provides based on students' responses to earlier questions (Yasuda et al., 2021). Combining a CD with computerized adaptive testing ensures accurate and efficient measurement of each student's proficiency level (Morphew et al., 2018) with the fewest questions necessary.

In this paper, we apply a CD model to student responses on the Calculus Concept Inventory (CCI, Epstein, 2006) and the Precalculus Concept Assessment (PCA, Carlson et al., 2010). Prior work establishing acceptable fit for the CD student model used in this study (Roberge et al., 2025). This application illustrates the additional resolution provided by skill-level proficiency estimates. We contrast pre–post changes in total scores with changes in skill proficiency profiles, including transition patterns across skills, to show what these two outcomes reveal and obscure.

Research Questions

To that end, we asked the following research questions for the PCA and CCI:

1. How much do students' scores change (learning) from pre to post on each of the instruments using mean shifts?
2. Using the CD student model, what are students' pre- and post-instruction skill-proficiency profiles, and to what extent does instruction increase proficiency probabilities and transitions into proficiency for calculus skills?

Theoretical Framework

Building on prior work (Vy et al, 2025), we drew on Evidence Centered Design (Mislevy et al., 2003) to develop the CCD. The conceptual assessment framework within Evidence-Centered Design has five components: student model, evidence model, task model, assembly model, and delivery system. In this article, we focus on the student and evidence models, as these components shape the assessment output. The student model includes the skills that the instrument aims to assess. The evidence model establishes the criteria for evaluating the assessment and defines how to score the student's answers to the questions. For the evidence models, we applied a CD model to interpret performance in terms of skill proficiency. The task model defines the format of the items (multiple-choice questions in our case). The assembly model integrates the student models with task models to assemble the assessment and assess student's performance. The delivery system (LASSO, Van Dusen et al, 2021) administers the assessment online.

Literature Review

Calculus can function as a gatekeeper in STEM education (Seymour et al., 2019). In fact, Seymour et al. (2019) found that calculus was one of two of "…the first and second year courses with the highest numbers of students receiving a D, F or incomplete/withdrawal grades…" at a rate of 20% (p. 199). Other studies have found DFWI rates as high as 34.3% (Koch & Drake, 2018). Calculus failure rates may disproportionately affect women (Ellis et al., 2016) and students of color (Treisman, 1992) contributing to a lack of representation in STEM fields (Sanabria & Penner, 2017). Evidence based instruction can greatly improve learning outcomes (Kramer et al., 2023) and such methods require evidence to assess. To this end, researchers have developed at least six RBAs to assess student readiness for calculus or learning during first and second semester calculus and provide data for researchers to evaluate.

Beginning in 2006, Epstein created the Calculus Concept Inventory (CCI) to measure learning in a first semester calculus course. In 2018 Thompson and Ashbrook began developing the Calculus 1 Concept Inventory (C1CI) and the Calculus 2 Concept Inventory (C2CI). More recently Lee et al. (2025) have released the Revised Calculus Concept Inventory (RCCI) which combined the CCI and C1CI. Carlson et al (2010) developed the Pre-calculus Concept Assessment (PCA) to measure student learning in precalculus. She and collaborators would later extend this work to the Calculus Concept Readiness (CCR) in (2015) to measure student proficiency on prerequisite concepts important for calculus.

Because researchers use a wide range of instruments and inconsistent reporting practices, we found it difficult to characterize learning gains in calculus. Researchers sometimes use RBAs such as the CCI (Bagley, 2014; Villalobos-Camargo, 2021; Dagley et al, 2018; Epstein, 2007 Thomas & Lozano 2013) but often they create their own instruments for pre and post assessments (Adams and Dove, 2018; Cronhjort et al., 2013; Kramer et al., 2023; Young et al, 2011). Reporting of descriptive statistics varies across these studies. Normalized gains in lecture based courses tend to be small and range from 0.1 (Villalobos-Camargo, 2021) to 0.3 (Young et al., 2011; Adams and Dove, 2018). Normalized gains (Hake et al., 1998) are not without issues (Nissen et al., 2018).

We (Roberge et al., 2025) identified a list of calculus skills by examining the learning outcomes from the OpenStax first semester calculus textbook (Strang & Herman, 2016) and from the Advanced Placement AB calculus test (*AP Calculus AB and BC: Course and Exam Description*, 2020). We initially developed a list of 27 skills that we used to code the items from both RBAs. Each instrument had too few items to cover 27 skills. After several iterations, we decided to focus on five broader categories: Prerequisite skills, Limits, Derivatives, Applications of Derivatives and Introduction to Integration. Tallman et al. (2016), after analyzing 254 different final exams for post-secondary, first-semester calculus courses, came to a similar conclusion about common calculus content. Our results (Roberge et al., 2025) supported the student models of the five skills fitting the data well. Both fit characteristics (RMSEA and SRMSR) were well below cutoff values and classification accuracies were either acceptable or good for each learning outcome.

Methods and Materials

The data was supplied by the LASSO platform (LASSO, 2023). The data consisted of 1,366 college student responses to both a pre and a posttest from the PCA (25 items – 897 students) and CCI (22 items – 469 students). Data for the PCA came from six institutions across thirty-three unique math courses that ran between the fall of 2016 and the summer of 2018. Data for the CCI came from five institutions, eight instructors, and eighty-six courses taught between the Fall of 2021 and the Winter session of 2024. Following previous data cleaning processes on LASSO data (Morley, Nissen & Van Dusen, 2023; Van Dusen & Nissen, 2020) we excluded students who took less than five minutes to complete the assessment or attempted fewer than 80% of the questions.

Descriptive statistics on pre/post scores for the PCA and CCI suggest only small changes. These scores are on a 100 point percentage scale. The PCA pre and post score means were 44.4/46.0 with standard deviations 17.1/17.3 respectively. The CCI pre and post score means were 34.2/38.7 with standard deviations 15.8/18.2 respectively. The PCA effect size was d = 0.09 and the CCI effect size was d = 0.26.

To determine if the differences in mean scores were significant, research question 1, we used a hierarchical linear model. Pre and post scores are nested within students and students are nested within courses. Neglecting nesting in data can violate the assumption of independence of many statistical methods such as multiple regression and ANOVA (Van Dusen & Nissen, 2019). For our models the intraclass correlation coefficients (ICCs) were, for PCA/CCI, students 0.54/0.56 and for courses was 0.18/0.069 respectively. Greater between group variance indicated we needed to treat groups (in this case: students and courses) separately and the hierarchical linear model accounts for both within-student variance and within-course variance.

To answer research question 2 we used CD models to produce finer grained information than traditional concept inventories. The model used test items, student responses, and a binary Q-matrix that codified item-skill dependencies. The Q-matrix is an $n \times m$ matrix for n items and m skills (Chen et al., 2015; Cui, Gierl & Chang, 2012). A one in the *ij-th* location means that skill j is required to correctly answer item i. An item may require more than one skill so its row in the Q-matrix will be a string of zeroes and ones. The resulting CD fitted to data categorized students into skill profiles and provided a variety of classification accuracies (Ravand & Robitzsch, 2015).

Skill profiles are binary vectors of length m (for m skills) where each '1' represents a proficiency and each '0' represents a missing proficiency. For example, a skill profile of 10100 for the CCI represents proficiency on prerequisite material and derivatives but missing proficiency on limits, applications of derivatives and introductory integration. We used alluvial plots to show how students transitioned from one skill profile to another from the pretest to the posttest. The quantity of students following any particular path from left to right is depicted by the size of the band connecting the left node to the right node. Additionally, when presenting alluvial plots of student proficiency groups we filtered to include only groups of ten or larger. For the PCA, this resulted in 24 students removed and six skill profiles removed. For the CCI this resulted in 39 students removed and fourteen skill profiles removed.

Table 1. The number of items and the classification accuracy (CA) for each skill for each RBA.

		Prerequisites	Limits	Derivatives	Apply Deriv.	Integration
PCA	items	20	3	0	5	1
	CA	0.92	0.68	-	0.87	0.91
CCI	items	6	3	6	6	1
	CA	0.94	0.91	0.94	0.95	0.98

Skill classification accuracy is a value between 0 and 1. Values between 0.8 and 0.9 are considered acceptable and values above 0.90 are considered high. For the five skill student model applied to the PCA and CCI data, Roberge et al. (2025) reported the results in Table 1. The reader may notice that the skill classification accuracy for Introductory Integration is considered high even though each RBA has only a single item coded for this skill. When a skill is represented by only one or two items, we may be skeptical of a high classification accuracy. On the PCA only 17% answered item 8 correctly on the pre-test (coded for integration) and only 18% on the post-test. On the CCI only 20% of students correctly answered item 18 (coded for integration) on the pre-test and only 24% on the post-test.

Findings

The results, shown in Table 2, indicate that post-test scores for both RBAs were larger than pre-test scores. The average post-test score on the PCA, accounting for student level and course level nesting of data, is 1.57 percentage points higher than the average pre-test score and this difference is statistically significant at the level of $p < 0.001$. The average post-test score on the CCI, similarly accounting for nested data, is 4.54 percentage points higher than the average post-test score and this difference is statistically significant at a level of $p < 0.001$.

Table 2. Model summary for each of the hierarchical linear models

		coefficient	st. error	df	t	p
PCA	intercept	39.31	1.74	26.84	22.56	$p < 0.001$
	post-test	1.57	0.45	897.00	3.52	$p < 0.001$
CCI	intercept	34.52	0.95	107.03	36.40	$p < 0.001$
	post-test	4.54	0.66	474.11	6.92	$p < 0.001$

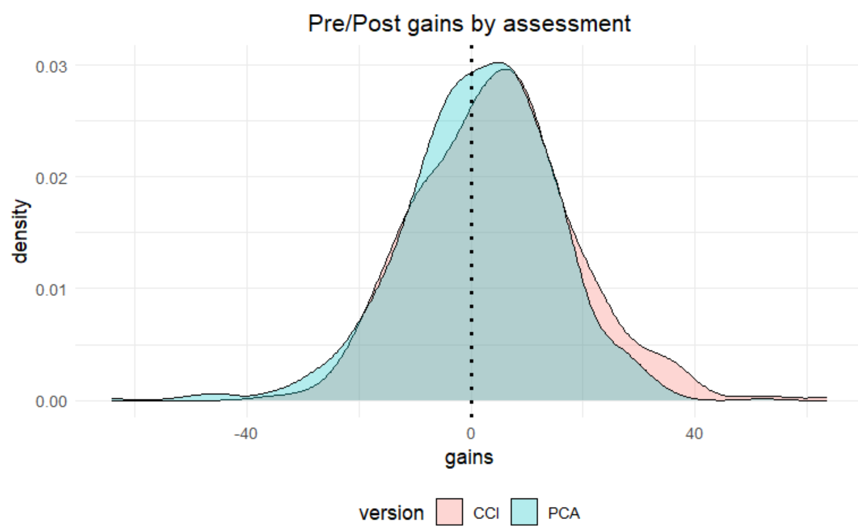

Figure 1. Density plots by gain (post score - pre score) for the CCI and PCA.

Addressing RQ2, Figure 2 depicts the changes in student skill profiles from pretest to posttest. The results for the PCA are located on the left and the results for the CCI are located on the right. Pretest skill profiles are depicted on the left-hand side of each alluvial plot and posttest skill profiles are depicted on the right-hand side of each plot. For example, in each alluvial plot the profiles 0000 or 00000 (we did not code any items from the PCA as assessing "Derivatives") would represent a lack of proficiency for all of the skills. Profiles 1111 or 11111 would represent proficiency in all the skills. The size of each rectangle represents the quantity of students in each group, for example in both the PCA and CCI the largest skill profile group on the pretest is 0000 or 00000.

The two figures portray overall similar trends of small changes in skills that were consistent with the results from research question 1. The PCA also differs because the size of the 0000 group decreased by more than half (51%) whereas for the CCI the 00000 group only decreased by 12%. Large chunks of the 0000, 1100, 1110, and 1111 groups on the pretest stayed at that same skill proficiency on the posttest. For the 0000, and 1100 groups, a notable number of students gained additional skill proficiency on the PCA posttest. This was much smaller for the transition from 1110 to 1111. Almost all of the students that ended with proficiency of all 4 skills, 1111, started there or with proficiency of the first three skills.

The shifts on the CCI were smaller than on the PCA and there were fewer skill proficiency profiles, which was likely due to the smaller sample size. We treat skill proficiency profiles of

00000 and 00001 as equivalent because that last skill is integration and the CCI only has one item that assesses integration. Only mastering integration is unlikely. When we equate skill profiles 00000 and 00001, out of the 378 students who start with no proficiency, 336 finish the course in the same situation. That's 89% of students acquiring no increased proficiency. Though enough students transitioned from 00000 to either 11110 or 11111 to make both of those groups increase in size. The total number of students who finished with proficiency in four or five of the skills was 17% (37) of the sample.

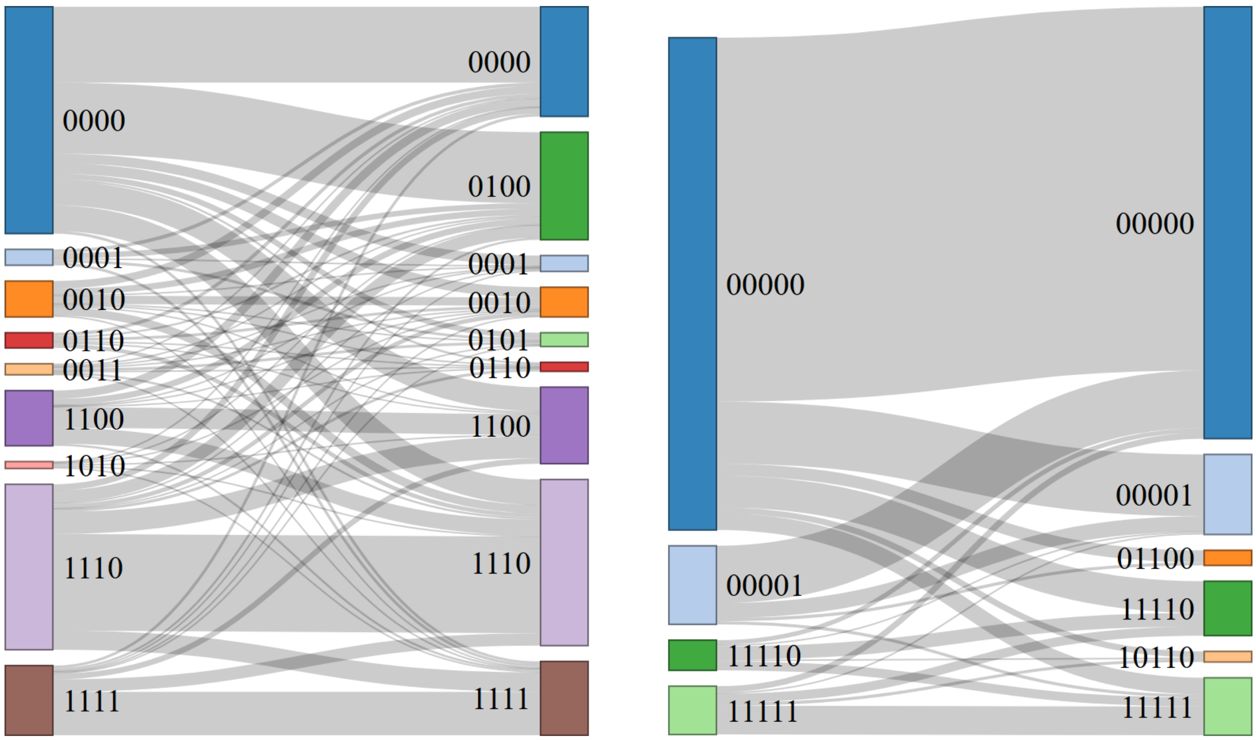

Figure 2. Sankey diagrams of pre-test and post-test data for the PCA (left) and CCI (right). Pre-test skill profiles are on the left of each plot and post-test skill profiles are on the right. For the PCA the binary 4-vector represents student proficiency on prerequisites, limits, applications of derivatives and introductory integration. For the CCI the binary 5-vector represents student proficiency on prerequisites, limits, derivatives, applications of derivatives and introductory integration. For example, on the CCI plot the skill profile 01100 would represent lacking proficiency in prerequisite concepts, proficient in limits and derivatives while lacking proficiency in applications of derivatives and introductory integration. Note: the data was filtered to include only skill profiles that contained at least ten students. Care should be taken when interpreting these plots as they do not represent individual courses but a population of students from many courses.

Limitations

These results represent the outcomes of students in the dataset and do not generalize to other courses or data. The contrast of skill proficiency profiles and overall scores, however, does generalize beyond this data. While binary skill proficiency profiles may seem to oversimplify student knowledge, they are a tradeoff of the finer grain measures that CDMs provide, and our development of the CCD seeks to better leverage this trait of CDMs. The timeframe of pre and post scores misses within-semester dynamics that CDAs can assess. Finally, neither RBA was designed as a cognitive diagnostic and the PCA was not designed with calculus content in mind though the retrofitting of a non-diagnostic assessment in this way is common practice (Ravand & Robitzsch, 2015) and serves to illustrate the information CDAs can provide.

Discussion

While both post-test scores were larger than the pre-test scores, the difference shows limited learning. The effect size on the PCA was d=0.09 and on the CCA was d=0.26. Kraft (2020) points out that a typical gain in secondary school is d=0.5 over an academic year and in introductory college physics the typical gain is closer to d=1.0 (Nissen et al., 2022). The density plots in Figure 1 rule out both floor and ceiling effects, which also indicates that the instruments did not grossly misalign with instruction. The results of the PCA and CCI data are consistent with existing literature on learning gains in calculus classrooms even though those gains are hard to compare given the variety of instruments. This points to a need to develop a standardized, yet flexible, instrument that can meet a broad range of teaching and research needs.

The CDM provides a finer grained perspective than the overall scores. Some students did make small gains in their skill proficiency. This is notable in the decrease of the no skill proficiency group from pre to posttest on the PCA and the increase in the 4 and 5 skill proficiency groups on the CCI. These small gains, however, sit within a larger picture of stagnation where instruction maintains proficiency for those who start with it and leaves those without proficiency in that state. It is particularly concerning that there does not seem to be a greater increase in the number of students gaining proficiency on prerequisite skills.

Consistent with the results from our first research question, either the instruction in these courses did not support the development of student proficiency or the RBA was not a good fit for the content of the course. Our five skills are consistent with commonly covered content areas in a first semester calculus course (Tallman et al. 2016) and most students who start with no proficiency (00000) end the course with no change to that skill profile. Most students who began with full proficiency retained that throughout the course, though some students transitioned from partial/no proficiency to full proficiency. This is consistent with Bressound's (2021) description of calculus classrooms in higher education having both students who are encountering these ideas for the first time and for whom this may be their last math course and those taking calculus again, after high school preparation who are building towards a math intensive STEM discipline.

Conclusion

The data for the PCA and CCI pre and post test scores are consistent with the literature that students experience relatively little learning in introductory calculus courses. Students who begin the course with low proficiency have minimal opportunity to improve while those who already possess high proficiency maintain that level.

The post-test score is too little too late to support teaching and learning. The CCD will support instruction by providing timely assessments that are fine grained enough to be acted upon. For example, instructors can use CCD results to inform what content they cover in their lectures or on their homework and to inform timely interventions to better ensure that students had the skills they needed before moving on. Student groups for in-class work could also be structured using CCD results to foster peer collaboration. The CCD, delivered with computerized adaptive technologies on the LASSO platform, has the capacity to support this vision.

References

AP Calculus AB and BC: Course and Exam Description. (2020, Fall). The College Board. https://apcentral.collegeboard.org/media/pdf/ap-calculus-ab-and-bc-course-and-exam-description.pdf

Adams, C., & Dove, A. (2018). Calculus students flipped out: The impact of flipped learning on calculus students' achievement and perceptions of learning. Primus, 28(6), 600-615.

Bagley, S. F. (2014). Improving student success in calculus: A comparison of four college calculus classes (Doctoral dissertation, San Diego State University).
Bressoud, D. M. (2021). The strange role of calculus in the United States. *ZDM–Mathematics Education*, *53*(3), 521-533.
Brewe, E. (2008). Modeling theory applied: Modeling Instruction in introductory physics. American Journal of physics, 76(12), 1155-1160.
Carlson, M., Oehrtman, M., & Engelke, N. (2010). The Precalculus Concept Assessment: A Tool for Assessing Students' Reasoning Abilities and Understandings. *Cognition and Instruction*, *28*(2), 113–145. http://www.jstor.org/stable/27806346
Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850-866.
Cronhjort, M., Filipsson, L., & Maria, W. (2013). Can Peer Instruction in calculus improve student learning?. In CDIO Conference. CDIO.
Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19-38.
Dagley, M. A., Gill, M., Saitta, E., Moore, B., Chini, J., & Li, X. (2018). Using active learning strategies in calculus to improve student learning and influence mathematics department cultural change. In Proceedings of the Interdisciplinary STEM Teaching and Learning Conference (2017-2019) (Vol. 2, No. 1, pp. 52-65).
Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences, 116(39), 19251-19257.
De La Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicología Educativa, 20(2), 89-97.
Ellis, J., Fosdick, B. K., & Rasmussen, C. (2016). Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PloS one*, *11*(7), e0157447.
Epstein, J. (2006). The calculus concept inventory. *National STEM Assessment, Washington, DC*, 60-67.
Epstein, J. (2007, September). Development and validation of the Calculus Concept Inventory. In Proceedings of the ninth international conference on mathematics education in a global community (Vol. 9, pp. 165-170). Charlotte, NC.
Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the national academy of sciences, 111(23), 8410-8415.
Gleason, J., Bagley, S., Thomas, M., Rice, L., & White, D. (2019). The calculus concept inventory: a psychometric analysis and implications for use. *International Journal of Mathematical Education in Science and Technology*, *50*(6), 825-838.
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*(1), 64-74.
Jones, B. L. (2021). *A Psychometric Analysis of the Precalculus Concept Assessment*. [Unpublished doctoral dissertation] Brigham Young University.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Koch, A. K., & Drake, B. (2018). Digging into the disciplines: The impact of gateway courses in accounting, calculus, and chemistry on student success. *Chemistry*, *34*, 29-4.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, *49*(4), 241-253.

Kramer, L., Fuller, E., Watson, C., Castillo, A., Oliva, P. D., & Potvin, G. (2023). Establishing a new standard of care for calculus using trials with randomized student allocation. *Science*, *381*(6661), 995-998.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

Learning about stem student outcomes (LASSO) (2023), URL https://lassoeducation.org.

Le, V., Nissen, J. M., Tang, X., Zhang, Y., Mehrabi, A., Morphew, J. W., ... & Van Dusen, B. (2025). Applying cognitive diagnostic models to mechanics concept inventories. Physical Review Physics Education Research, 21(1), 010103.

Roberge, K., Le, V., Van Dusen, B., Nissen, J. M., (2025) Calculus Cognitive Diagnostic: Mathematics skills tested in first semester calculus courses. In *2025 Physics Education Research Conference Proceedings* (pp. XX-YY).

Lee, J., Yi, M., Song, W., Dwyer, J., Williams, B., & Moskal, B. (2025). Revised Calculus Concept Inventory: Measuring the Learning Gains of Students in a First Calculus Course. *International Journal of Research in Undergraduate Mathematics Education*, 1-22.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Morley, A., Nissen, J. M., & Van Dusen, B. (2023). Measurement invariance across race and gender for the Force Concept Inventory. *Physical Review Physics Education Research*, *19*(2), 020102.

Morphew, J. W., Mestre, J. P., Kang, H. A., Chang, H. H., & Fabry, G. (2018). Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course. *Physical Review Physics Education Research*, *14*(2), 020110.

Nissen, J. M., Talbot, R. M., Nasim Thompson, A., & Van Dusen, B. (2018). Comparison of normalized gain and Cohen'sd for analyzing gains on concept inventories. Physical Review Physics Education Research, 14(1), 010115.

Nissen, J. M., Many Horses, I. H., Dusen, B. V., Jariwala, M., & Close, E. (2022). Providing context for identifying effective introductory mechanics courses. The Physics Teacher, 60(3), 179-182.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. Practical Assessment, Research, and Evaluation, 20(1).

Roberge, K., Lee, V., Van Dusen, B., Nissen, J. M., (2025) Calculus Cognitive Diagnostic: Mathematics skills tested in first semester calculus courses. *Proceedings of Physics Education Research Conference*. (2025)

Sanabria, T., & Penner, A. (2017). Weeded out? Gendered responses to failing calculus. *Social Sciences*, *6*(2), 47.

Seymour, E., Hunter, A. B., & Weston, T. J. (2019). Why we are still talking about leaving. In *Talking about leaving revisited: Persistence, relocation, and loss in undergraduate STEM education* (pp. 1-53). Cham: Springer International Publishing.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. Science, 359(6383), 1468-1470.

Strang, G., & Herman, E., (2016). *Calculus Volume I*. OpenStax

Tallman, M. A., Reed, Z., Oehrtman, M., & Carlson, M. P. (2021). What meanings are assessed in collegiate calculus in the United States?. *ZDM–Mathematics Education*, *53*(3), 577-589.

Thomas, M., & Lozano, G. (2013). Analyzing Calculus Concept Inventory gains in introductory calculus. In *Proceedings of the Sixteenth Annual Conference on Research in Undergraduate Mathematics Education.-Denveepstein2006calculusr, CO: SIGMAA* (Vol. 2, pp. 637-646).

Thompson, P. W., Milner, F., & Ashbrook, M. (2018). Final report for DUE-1625678: Project DIRACC: Developing and investigating a rigorous approach to conceptual calculus. National Science Foundation. http://patthompson.net/ThompsonCalc/AnnualReports.pdf

Treisman, U. (1992). Studying students studying calculus: A look at the lives of minority mathematics students in college. *The College Mathematics Journal*, *23*(5), 362-372.

Van Dusen, B., & Nissen, J. (2019). Modernizing use of regression models in physics education research: A review of hierarchical linear modeling. Physical Review Physics Education Research, 15(2), 020108.

Van Dusen, B., & Nissen, J. (2020). Equity in college physics student learning: A critical quantitative intersectionality investigation. *Journal of Research in Science Teaching*, *57*(1), 33-57.

Van Dusen, B., Shultz, M., Nissen, J. M., Wilcox, B. R., Holmes, N. G., Jariwala, M., ... & Pollock, S. (2021). Online administration of research-based assessments. *American Journal of Physics*, *89*(1), 7-8.

Villalobos-Camargo, G. (2021). The Calculus Concept Inventory Applied to the Case of Large Groups of Differential Calculus in the Context of the Program" Ser Pilo Paga" in Colombia. Mathematics Teaching Research Journal, 13(3), 208-223.

Yasuda, J. I., Mae, N., Hull, M. M., & Taniguchi, M. A. (2021). Optimizing the length of computerized adaptive testing for the Force Concept Inventory. *Physical review physics education research*, *17*(1), 010115.

Young, C. Y., Georgiopoulos, M., Hagen, S. C., Geiger, C. L., Dagley-Falls, M. A., Islas, A. L., ... & Bradbury, E. E. (2011). Improving student learning in calculus through applications. International Journal of Mathematical Education in Science and Technology, 42(5), 591-604.