Calculus Cognitive Diagnostic: Mathematics skills tested in first semester calculus courses

Kevin Roberge
University of Maine, Orono, ME, 04473, USA

Vy Le and Ben Van Dusen
School of Education, Iowa State University, Ames, IA, 50011, USA

Jayson M. Nissen
Department of Physics, Montana State Universiy, Bozeman, MT, 59715, USA

Physics instructors and education researchers use research-based assessments (RBAs) to evaluate students' preparation for physics courses. This preparation and learning can cover a range of topics, including algebra, pre-calculus and calculus. To meet the need for high quality, adaptable assessments, we are developing a cognitive diagnostic for calculus as an online test using Evidence Centered Design. Using cognitive diagnostic models and computerized adaptive testing allows the users to customize which skills are assessed and when they are assessed. In this article, we investigate the extent to which our model of the students calculus skills fit the data from three mathematics RBAs. Our model included five calculus skills students often use in physics courses: pre-calculus knowledge, limits, derivatives, applications of derivatives, and integration. The data came from the LASSO platform and includes 4,132 responses from the Calculus Concept Assessment, Calculus Concept Inventory, and Pre-calculus Concept Assessment. The "Deterministic Input, Noisy 'And' gate" (DINA) analyses demonstrated a good fit for the five skills to the data. These results provide validity for our model of mathematics skills students develop. Including items from these RBAs will provide a foundation for developing a flexible assessment of a range of mathematics skills that can adapt to instructors' and researchers' needs.

2025 PERC Proceedings edited by Pawl, Zwolak, and Leak; Peer-reviewed, doi.org/10.1119/perc.2025.pr.Roberge

I. INTRODUCTION

Using existing research based assessments (RBAs) to assess relevant mathematics skills poses physics researchers with the dilemma of choosing what to measure. All RBAs on PhysPort, for example, use fixed-length designs that require using the whole assessment or risking undermining the validity arguments for the assessment. They also tend to focus on an overall ability measure that is represented by their total score, rather than providing specific construct or skill measures. This inflexible design and length limits the ability to use these assessments flexibly and effectively, such as for weekly quizzes during the semester to study the impact of an intervention. It also limits the quality of the data produced and the research that can be conducted. To solve these problems we are developing a calculus cognitive diagnostic (CCD) delivered online using computerized adaptive testing. Cognitive diagnostic models measure student understanding of multiple skills as opposed to a single skill or attribute measured by a traditional RBA [1]. Computerized adaptive testing addresses this issue by optimizing the information each question provides based on students' responses to earlier questions [2]. This design allows the assessment to accurately measure each student's proficiency level [3] with the fewest questions necessary.

Physics education researchers have investigated how students use calculus to reason about physical phenomena across the undergraduate curriculum. Foundational concepts such as integration and differentials underpin more advanced reasoning in areas like thermodynamics [4], electromagnetism [5], and quantum mechanics [6]. Students struggle to apply these tools in physics, despite procedural fluency gained in mathematics courses [7, 8]. Research on upper-division topics reveals that difficulties with partial derivatives [4], Taylor series [9], and coordinate-based volume elements [10, 11] can be traced back to limited understanding of basic calculus ideas applied in novel contexts. For example, Meredith and Marrongelle [12] found that an insufficient understanding of limits may impede student use of integration in a problem solving capacity. Although, conversely, Nguyen and Rebello [8] found that students may recognize the need for an integral and be able to integrate but be unable to setup the integral in an electrostatics context. These are nuanced observations that require granular assessment.

When the CCD is complete it will assess 27 skills–from across five broader skill sets: prerequisites, limits, derivatives, applications of derivatives, and integration–taught within most first semester calculus courses and used across a range of physics courses. These 27 skills would include 5 prerequisite mathematics skills found in a pre-calculus course. The CCD could be used as a pre and post term, monthly, or weekly as a formative assessment over modules or fine grained skills or as a combination of all of these. For example, not only could specific prerequisite knowledge such as trigonometric functions be assessed but so could conceptual understandings of derivatives and graphs. Using Evidence-Centered Design [13] we have developed a student model of skills necessary for a first semester calculus course.

II. RESEARCH QUESTIONS

In this first step in the development of the CCD, we focus on the validity and usefulness of the student model. If the data fits our student model well, then we can move forward with using this model for the CCD and combine the items from the RBAs to populate the initial item bank. To that end, we asked the following research questions for three mathematics RBAs: the Calculus Concept Assessment (CCA) [14], the Calculus Concept Inventory (CCI) [15] and the Pre-calculus Concept Assessment (PCA) [16].

- How well does the data from the three RBAs support the student model of the skills we propose to measure with the CCD?
- How well do each of the three RBAs measure students' mastery of the skills in the student model?

To support the readers' interpretation of our research, Table III includes a selection of terms and their definitions [17].

III. THEORETICAL FRAMEWORK

Building on previous work [17], we drew on Evidence-Centered Design [13] to develop the CCD assessment. The conceptual assessment framework within Evidence-Centered Design has five components: student models, evidence models, task model, assembly model, and delivery system. In this article, we focus on the student and evidence models, as these components align with our research objectives. The student model consists of 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 cognitive diagnostic model to interpret student performance in terms of mastery levels (more details are in Sec. V of the work [17]). 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 estimate overall student's performance. The delivery system, where we used the LASSO platform [18, 19], administers the assessment online.

IV. METHODS

The LASSO platform provided the data for this research [18]. The data consisted of 4,168 college student responses across three post-course assessments: the CCA (17 items - 1,463 students), CCI (22 items - 1409 students), and PCA (25 items - 1,260 students). Data for the CCI came from five institutions, ten instructors over 123 courses (121 mathematics and 2 physics) taught between the fall of 2021 and the Winter session of 2024. Data for the CCA came from seven institutions, twelve instructors over 42 courses (35 mathematics and 7 physics) taught between the spring of 2019 and the fall of 2023. Data for the PCA came from eight institutions across forty unique math courses that ran between the fall of 2015 and the summer of 2018. The analysis excluded students who took less than five minutes to complete the assessment or attempted less than 80% of the questions. In cases where students completed the same assessment multiple times, we only used their first response to the most recent post test. Descriptive statisics are provided in Table I

TABLE I. RBA descriptive statistics (percentage points)

	n	mean	median	standard deviation
CCA	1,463	37.7	33.3	17.5
CCI	1,409	38.5	36.4	18.1
PCA	1,260	44.3	42.3	17.7

The research employed an iterative, mixed-methods approach. The research team has experience teaching precalculus through differential equations as well as calculus-based physics courses. First, we identified a set of skills. We then coded the items from the RBAs for each of those skills. An item could measure multiple skills. This coding produced a Q-matrix [20, 21], which links each item to each skill. For an assessment with n items measuring k skills, the Q-matrix is a $n \times k$ matrix of ones and zeros. A one in the ijth position means that skill j is required to answer item i. In the next step, we used "Deterministic Input, Noisy 'And' gate" (DINA) models to test the fit of each Q-matrix to the data from each RBA. Fitting the model produces suggested changes to the Q-matrix. We reviewed the suggestions and determined whether or not to implement them. We then tested each updated Q-matrix that integrated our review of the suggested changes and reevaluated model fit using DINA.

We identified the list of skills by examining the learning outcomes from the OpenStax first semester calculus textbook [22] and the learning outcomes from the Advanced Placement AB calculus test [23]. We developed a list of 27 skills that fell into five broader categories, and coded the three RBAs with those skills. Initially the resulting Q-matrices were too sparse; for most of the skills we lacked any questions or even enough questions to estimate parameters. Of the 27 learning outcomes, eight lacked any item coverage and total of 17 contained less than two items. After several iterations, we decided to focus on the five broader categories (see Table II) that represent common themes in many first semester calculus courses. The questions for the CCI, CCA and PCA were independently coded by two of the authors who would then discuss and reconcile differences with each other and the third author. These were each evaluated by a DINA model, which produced suggested changes to the coding.

TABLE II. Names and Descriptions of current skills

Skills	Description/example content
Prerequisite skills	Content of a pre-calculus course including an understanding of linear, power, polynomial, trigonometric, exponential, and logarithmic functions.
Limits	An understanding of limits and limiting processes. Basic limit laws, limits from both tables of data and graphs of functions.
Derivatives	The meaning of the tangent line. Determining the slope of the tangent line from calculation, a graph or a table. Comparing instantaneous and average rates. The relationship between graphs of y' and y. Higher derivatives.
Applications of Derivatives	Problems with extrema. Optimization problems. Units of derivatives. Using the first and second derivative to solve problems.
Integration	Understanding integration as accumulation, as the area between curves. Basic integration properties, units on integrals and the Fundamental Theorem of Calculus.

DINA (Deterministic Input, Noisy 'And' gate) models are a type of cognitive diagnostic model that classifies learners based on which skills they have mastered [32]. An assumption of DINA models is that a student must have all the required skills to correctly answer an item on an assessment. That an item requires certain skills is encoded in the Q-matrix. The probability of a student answering an item correct even when they do not have all the required skills is called guessing and this is estimated within the model. The probability that a student answers an item incorrectly even when they have those required skills is called slipping and is also estimated within the model. DINA estimates the probability that a student with certain skills will answer a particular item depending on the skills required for that item along with the guess and slip estimates for that item.

To fit a DINA model to our data and Q-matrices we used the G-DINA package in R [34]. Two fit indices were calculated to assess the DINA model fit. The root mean square error of approximation (RMSEA), which is based upon the M2 statistic, and the standardized root mean square residual (SRMR). For both of these indices, values less than 0.05 indicate a good fit while values between 0.05 and 0.08 are considered an acceptable fit. [35–37]. G-DINA also calculated the classification accuracy scores to quantify the accuracy of the model's estimation of the proportion of students whose proficiency status for a skill was correctly determined by the model. Classification accuracy values are between 0-1 and values greater than or equal to 0.9 are interpreted as high [38, 39] while values between 0.8 and 0.9 are considered ac-

TABLE III. Definitions of terms.

Term - Definition
Computerized adaptive testing (CAT) - Administered on computers, the test adaptively selects subsequent test items based on each test taker's previous responses to match the demonstrated proficiency [3, 24, 25].
Proficiency - "...the student's general facility with answering the items correctly on the assessment under consideration". Higher proficiency increases the probability of answering assessment items correctly. Different fields use different terms for proficiency, such as skill, ability, latent trait, and omega.
Skills - A latent attribute that students need to master to answer items correctly and that cuts across content areas [24, 26, 27].
Q-matrix - A Q-matrix, or "question matrix," is a binary matrix that maps the relationship between test items and the underlying skills they measure [20, 21].
Cognitive diagnostic (CD) assessment - An assessment method that evaluates students on specific skills to determine mastery. In contrast to traditional assessment methods that measure students on a single proficiency, CD provides diagnostic information on skill strengths and weaknesses to support personalized educational strategies [28, 29].
Classification accuracy - The agreement between observed and true skill classifications. In practice, this is calculated using the expected skill classifications rather than the true classifications, which is detailed in an example around Equations 4 and 6 in Ref. [30].
Deterministic Input, Noisy 'And' gate (DINA) model - A cognitive diagnostic model assuming that a student must master all the required skills to solve an item correctly. The absence of any required skills cannot be compensated by the mastery of others. This model operates within a binary framework, categorizing each skill as either mastered or not mastered [29, 31–33].
Evidence-centered design (ECD) model - A framework for developing educational assessments based on establishing logical, evidence-based arguments [13].

ceptable [40].

V. FINDINGS

The final Q-matrices are summarized in Table IV. The CCA and CCI assessed all five skills, while the PCA did not assess derivatives. The sum of the item counts for each RBA in Table IV exceeds the total number of items on each RBA because five items across the three RBAs assessed more than one skill. As the name implies, the PCA primarily assessed precalculus skills whereas the CCA and CCI had more balance in the number of items assessing each skill. Many of the skills were only assessed by 1 to 3 items. Integration, which lies on the edge of content covered in first semester calculus courses, had the fewest number of items assessing it across all three instruments.

In answering RQ1, two results support the student models of the five skills fitting the data well. First, across the three RBAs, the initial DINA model suggested 24 out of 320 possible changes. The coding team adopted three of those suggestions, resulting in an overall change rate of 1%. This included one change to the Q-matrix for the CCA and two changes to the Q-matrix for the CCI. The second result supporting the student model was the DINA model fit. The DINA models for the final coding for each of the three RBAs demonstrated good fit with RMSEA less than 0.04 and SRMSR less than 0.05 for all three assessments (see Table V).

TABLE IV. Distribution of items by RBA and skills

	CCA	**CCI**	**PCA**	**Total**
Prerequisites	7	6	20	33
Limits	4	3	3	10
Derivatives	3	6	0	9
Applications of Derivatives	2	6	5	13
Integration	2	1	1	4

TABLE V. DINA model fit statistics

	CCA	CCI	PCA	Cutoff
RMSEA	0.035	0.018	0.034	0.050
SRMSR	0.037	0.034	0.043	0.050

The DINA models output 14 classification accuracies across the three RBAs (the PCA did not assess derivatives) that are shown in Table VI. Two of the 14 classification accuracies are below 0.8, which is our lower acceptable boundary for low-stakes assessments. Nine of the 14 classification boundaries were above 0.9, which was our lower acceptable boundary for high-stakes assessments. None of the RBAs had consistently low classification accuracies. The CCI had classification accuracies greater than 0.9 for all five skills. Nor were the classification accuracies consistently poor for any skill across all three RBAs. However, some care needs to be taken when interpreting classification accuracies. For example, the classification accuracy was 0.87 or higher for integration on all three RBAs though it was measured by one to two items per assessment. We examine this further in the Discussion section.

TABLE VI. Classification Accuracies by RBA and skill

	CCA	**CCI**	**PCA**
Prerequisites	0.90	0.94	0.92
Limits	0.90	0.91	0.68
Derivatives	0.88	0.94	-
Applications of Derivatives	0.78	0.95	0.87
Integration	0.87	0.98	0.91

VI. DISCUSSION AND LIMITATIONS

Calculus skills are foundational for success in physics. Understanding how and when students are (or are not) able to transfer those skills across disciplinary boundaries can improve the impact of physics courses on student learning. The model fit statistics confirm that the items from the three RBAs for mathematics align well with our student model of the five skills. Each DINA model fit well, indicating that the items from each RBA can assess the skills in the student model. These results indicate that researchers could use these instruments to investigate more specific skills than overall calculus ability. The results also indicate that our proposed student model will work well for building the CCD. Because the student model fit all three RBAs well, combining their items should provide a adequate initial item bank for the CCD.

The three RBAs can each assess at least two skills with high classification accuracies and all fives skills were assessed with high classification accuracy by at least one instrument. Combining the items from these instruments into a single item bank for a cognitive diagnostic to assess these five skills should produce high classification accuracies for each of the skills. We are, however, concerned about the high classification accuracy for integration on each instrument with so few items. After an extensive literature review, we are not aware of any firm recommendations on the number of items to assess a skill. High classification accuracies with few items are more likely to occur for rare skill classifications, e.g., less than 17% of students across all three RBAs correctly answered the questions coded as integration. If the model also classified these students as non-proficient in the Integration skill that would result in a high classification accuracy. Because the CCD will serve a broad range of courses, the CCD will likely need more items to assess integration with a high classification accuracy than the four items from these three RBAs.

The findings may be limited by the data, which was almost exclusively from mathematics courses. Additionally, several of the skills do not yet have an adequate number of items. The analyses in this paper also do not reflect how the CCD will operate. Instead, they provide validity evidence for the student models for the CCD and to guide the development of the item bank for the CCD. Subsequent analyses of the CCD will investigate the extent to which the CCD measures the proposed student models of skills.

VII. CONCLUSIONS

The current work lays the foundation for the development of the Calculus Cognitive Diagnostic (CCD). By applying this model to data from three research-based assessments (RBAs), this study provides validity evidence for our student model for the CCD. Together, these contributions support the design of a system that can diagnose students' calculus understanding with precision and inform both instruction and research.

The CCD will allow investigating how specific calculus skills support student learning in physics. By targeting individual skills, the CCD will enable researchers and instructors to identify whether difficulties in physics courses—such as mechanics, thermodynamics, or quantum mechanics stem from misunderstandings in physics or from calculus. This precision will offer new insights into the role of mathematics in physics learning and support more effective instructional responses.

The CCD will offer flexibility in both content and timing of assessment. Instructors and researchers will be able to select which skills to assess and when, allowing for use as a pre/post assessment, a monthly test preparation tool, or a weekly formative check. Researchers can evaluate interventions with flexible timing and content coverage that can be reproduced across studies. The adaptive delivery will minimize testing time while maintaining measurement accuracy. Skill mastery profiles generated by the CCD will support individualized instruction and peer learning strategies—for example, by identifying students with complementary skill sets for group work. As the CCD item bank grows, it will support increasingly fine-grained assessments aligned with a wide range of instructional goals. The student model provides the foundation for this system and will ultimately support both classroom decision-making and research on student learning trajectories across mathematics and physics.

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

[2] J.-i. Yasuda, N. Mae, M. M. Hull, and M.-a. Taniguchi, Physical review physics education research **17**, 010115 (2021).

[3] J. W. Morphew, J. P. Mestre, H.-A. Kang, H.-H. Chang, and G. Fabry, Physical Review Physics Education Research **14**, 020110 (2018).

[4] R. R. Bajracharya, P. J. Emigh, and C. A. Manogue, Physical Review Physics Education Research **15**, 020124 (2019).

[5] R. E. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, Physical Review Special Topics—Physics Education Research **8**, 010111 (2012).

[6] M. Wawro, K. Watson, and W. Christensen, Physical Review Physics Education Research **16**, 020112 (2020).

[7] D. Hu and N. S. Rebello, Physical Review Special Topics—Physics Education Research **9**, 020108 (2013).

[8] D.-H. Nguyen and N. S. Rebello, Physical Review Special Topics—Physics Education Research **7**, 010113 (2011).

[9] T. I. Smith, J. R. Thompson, and D. B. Mountcastle, Physical Review Special Topics—Physics Education Research **9**, 020110 (2013).

[10] B. P. Schermerhorn and J. R. Thompson, Physical Review Physics Education Research **15**, 010111 (2019).

[11] B. P. Schermerhorn and J. R. Thompson, Physical Review Physics Education Research **15**, 010112 (2019).

[12] D. C. Meredith and K. A. Marrongelle, American Journal of Physics **76**, 570 (2008).

[13] R. J. Mislevy, R. G. Almond, and J. F. Lukas, ETS Research Report Series **2003**, i (2003).

[14] *Calculus concept assessment* (n.d.), URL https://lassoeducation.org/calculus-concept-assessment/.

[15] J. Epstein, in *Proceedings of the ninth international conference on mathematics education in a global community* (Citeseer, 2007), vol. 9, pp. 165–170.

[16] M. Carlson, M. Oehrtman, and N. Engelke, Cognition and Instruction **28**, 113 (2010).

[17] V. Le, J. M. Nissen, X. Tang, Y. Zhang, A. Mehrabi, J. W. Morphew, H. H. Chang, and B. Van Dusen, Physical Review Physics Education Research **21**, 010103 (2025).

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

[19] B. Van Dusen, M. Shultz, J. M. Nissen, B. R. Wilcox, N. Holmes, M. Jariwala, E. W. Close, H. Lewandowski, and S. Pollock, American Journal of Physics **89**, 7 (2021).

[20] Y. Chen, J. Liu, G. Xu, and Z. Ying, Journal of the American Statistical Association **110**, 850 (2015).

[21] Y. Cui, M. J. Gierl, and H.-H. Chang, Journal of Educational Measurement **49**, 19 (2012).

[22] G. Strang and E. Herman, *Calculus Volume 1* (OpenStax, 2016).

[23] C. Board, *AP calculus AB and BC course and exam description* (Fall 2020).

[24] H.-H. Chang, Psychometrika **80**, 1 (2015).

[25] D. J. Weiss, Applied psychological measurement **6**, 473 (1982).

[26] C. Helm, J. Warwas, and H. Schirmer, Empirical Research in Vocational Education and Training **14**, 1 (2022).

[27] T. Li and A. Traynor, AERA Open **8**, 23328584221081256 (2022).

[28] H. Ravand and A. Robitzsch, Practical Assessment, Research, and Evaluation **20**, 11 (2015).

[29] J. De La Torre and N. Minchen, Psicología Educativa **20**, 89 (2014).

[30] W. Wang, L. Song, P. Chen, Y. Meng, and S. Ding, Journal of Educational Measurement **52**, 457 (2015).

[31] E. Haertel, Applied Psychological Measurement **8**, 333 (1984).

[32] B. W. Junker and K. Sijtsma, Applied Psychological Measurement **25**, 258 (2001).

[33] J. de la Torre, Journal of Educational and Behavioral Statistics **34**, 115 (2009), ISSN 2193-1801.

[34] W. Ma and J. de la Torre, Journal of Statistical Software **93**, 1 (2020).

[35] D. Hooper, J. Coughlan, and M. Mullen, in *7th European Conference on research methodology for business and management studies* (2008), vol. 2008, pp. 195–200.

[36] L.-t. Hu and P. M. Bentler, Structural equation modeling: a multidisciplinary journal **6**, 1 (1999).

[37] S. T. Jang, American Educational Research Journal **55**, 1268 (2018).

[38] Z. Tan, J. De la Torre, W. Ma, D. Huh, M. E. Larimer, and E.-Y. Mun, Prevention Science **24**, 480 (2023).

[39] Q. Liang, J. de la Torre, M. E. Larimer, and E.-Y. Mun, Mendeley Data (2023).

[40] J. Paulsen, D. Svetina, Y. Feng, and M. Valdivia, Applied Psychological Measurement **44**, 267 (2020).