AI-Based Concept Inventories: Using Cognitive Diagnostic Computer Adaptive Testing in LASSO for Classroom Assessment

Dr. Jason Morphew, Purdue University

Jason W. Morphew is an Assistant Professor in the School of Engineering Education at Purdue University. He earned a B.S. in Science Education from the University of Nebraska and spent 11 years teaching math and science at the middle school, high school, and community college level. He earned a M.A. in Educational Psychology from Wichita State and a Ph.D. from the University of Illinois Urbana-Champaign.

Amirreza Mehrabi, Purdue Engineering Education

I am Amirreza Mehrabi, a Ph.D. student in Engineering Education at Purdue University, West Lafayette. Now I am working in computer adaptive testing (CAT) enhancement with AI and analyzing big data with machine learning (ML) under Prof. J. W. Morphew at the ENE department. My master's was in engineering education at UNESCO chair on Engineering Education at the University of Tehran. I pursue Human adaptation to technology and modeling human behavior(with machine learning and cognitive research). My background is in Industrial Engineering (B.Sc. at the Sharif University of Technology and "Gold medal" of Industrial Engineering Olympiad (Iran-2021- the highest-level prize in Iran)). Now I am working as a researcher in the Erasmus project, which is funded by European Unions (1M $_European Union & 7 Iranian Universities) which focus on TEL and students as well as professors' adoption of technology(modern Education technology). Moreover, I cooperated with Dr. Taheri to write the "R application in Engineering statistics" (an attachment of his new book "Engineering probability and statistics.")

Ben Van Dusen, Iowa State University of Science and Technology
Jayson Nissen

Computer Adaptive Testing in LASSO platform for classroom assessment and self-assessment

Jason W. Morphew[1], Amirreza Mehrabi[1], Ben Van Dusen[2], Jayson Nissen[3], & Hua Hua Chang[4]

[1] School of Engineering Education, Purdue University, West Lafayette, IN
[2] School of Education, Iowa State University, Ames, IA
[3] Nissen Education and Research Design, Monterey, CA
[4] College of Education, Purdue University, West Lafayette, IN

Abstract

Computerized Adaptive Testing (CAT) is a modern approach to educational technology that can transform classroom assessment and self-assessment strategies. CAT selects questions based on ability, item difficulty, and item discrimination at the moment which significantly reduces testing time. So, by considering measurement error, CAT ensures assessment accuracy, revealing a student's true ability level. The design of CAT within the Learning About STEM Student Outcomes (LASSO) platform adheres to a comprehensive spectrum of skills and attributes outlined by educators nationwide. CAT within the LASSO adeptly tailors question selection for each class, furnishing students with specialized reports grounded in distinct content. LASSO serves as a centralized platform enabling classes nationwide to access a diverse array of assessment contents and questions aligning with established educational standards, promoting frequent assessment. The amalgamation of CAT with cognitive diagnosis models within the LASSO platform empowers educators to gauge student mastery levels and confidently navigate the subsequent stages of the teaching process. Therefore, teachers can assess the effectiveness of their teaching methodologies, a vital aspect of their self-assessment.

Introduction

The emergence of artificial intelligence (AI) is driving a paradigm shift across education, particularly within STEM fields such as Physics and Engineering. The emergence of generative AI, large language models, and machine learning provides new and more powerful mechanisms for individualized and personalized learning. However, to realize the promise of AI in providing personalized learning, we must rethink assessment within introductory STEM courses by moving from static to adaptive assessments. Traditional assessment methods, while foundational, are often rigid and uniform. The static nature of these traditional exams limits their ability to conduct individualized assessments, failing to adequately assess skill and content mastery for diverse learners of all ability levels, leading to potential misrepresentations of the true abilities of students [1,2]. Furthermore, the static nature of these assessments can impact student motivation that stem from assessment that fail to adapt to individual student performance frustrating low-performing students while boring high-performing students [3]. In this context, Computerized Adaptive Testing (CAT) emerges as a transformative solution. Grounded in Item Response Theory (IRT), CAT dynamically adjusts question difficulty based on examinee responses. The ability to identify questions aligned with an individual's proficiency level allows for faster and more precise proficiency estimates, while providing more accurate measurement of students' conceptual understanding and skill mastery [4-7]. This ability to adaptively assess individual students can

provide more accurate and detailed information to tailor individualized interventions following assessment, while also improving test security [8].

This paper describes the development of the Mechanics Cognitive Diagnostic (MCD), a cognitive diagnostic computerized adaptive testing (CD-CAT) program that is hosted on the web-based Learning About STEM Student Outcomes (LASSO) platform [9]. This ongoing project aims to harness the potential of CAT on a national scale, creating an accessible and reliable assessment system for assessing conceptual STEM understanding for colleges and universities that aligns with STEM curriculum and uses Artificial Intelligence (AI) based assessment methods.

Table 1: Operational Definition of Terms

Term	Operational Definition	Example(s)
Proficiency	The proficiency of a person reflects the probability of answering test items correctly. The higher the individual's proficiency, the higher the probability of a correct response. Different fields refer to proficiency as ability, latent trait, theta.	• Percentage correct on static exams. • Theta estimate on CATs.
Content Area	Sub-divisions of course material that reflect distinct clusters or groupings of facts, concepts, theories and principles as defined by content experts.	• Conservation of Energy • 2D Kinematics
Concept	An idea that reflects the relationship between objects, fields, and forces that can explain natural processes or observations. Student conceptual understanding can be aligned with canonical explanations or can represent an alternative explanation (often referred to as misconceptions).	• An object in a state of motion remains in that state of motion unless acted upon by an outside force.
Skill	A procedural or conceptual operation that can be applied across content to answer assessment questions. Skills are assumed to be latent attributes that students need to master to correctly answer questions.	• Construct Force Diagrams • Interpret Graphs • Solve Trig Equations • Solve Questions Using Vectors

Literature Review

Traditional Assessments. Traditional assessments are static in question order, question difficulty, and exam length. This one-size-fits-all approach assumes that every question can equitably assess a wide diversity of student proficiencies. By providing static question difficulty, traditional assessments can potentially neglect student diversity in proficiency, potentially leading to misrepresentations of the students' actual skill mastery or conceptual understanding [10-14]. Complicating matters is that question difficulty is determined by a complex interaction between the content area, concepts, and skills being assessed. Traditional assessments typically do not distinguish between these levels, inadvertently linking skill mastery with concept and content difficulty, and non-content related skills such as reading level or familiarity with the question context. Traditional assessments that merely report a percentage correct also provide inefficient feedback for students by not providing students with information about concept or skill mastery

that would help direct their studying. Furthermore, standardized tests may inadvertently perpetuate biases, favoring certain demographic groups and contributing to disparities in educational outcomes based on factors such as race, class, and gender [15]. Finally, fixed question difficulty can create challenges for students who find tests excessively difficult or overly simplistic, which can impact motivation and course persistence. While static assessments have a long history within STEM education and can be designed well, most instructors have little or no training in psychometrics or writing assessment questions [16-19].

Because traditional assessments are static where all students receive the same questions, issues of academic integrity, test security, and equity remain a large problem. The ability for students to share test answers or post test questions to discussion boards makes reusing tests between sections or across years difficult, especially with the rise in computer-based and online tests [20-22]. While cheating concerns have been addressed by developing multiple forms or randomizing question order [23,24], these methods require extensive testing with large numbers of students to ensure equity of these parallel questions and large databases of questions in order to generate multiple forms of exams. Recently, randomized testing has emerged as a hybrid between traditional exams and CAT [25,26]. The randomized question selection allows students to leverage the testing effect by retaking exams while also increasing test security [27,28]. However, the extensive time and expense for developing the large test databases needed for equitable randomized exams are the same as for parallel forms above. In addition, randomized exams still typically only report percentage correct and identify which questions student get correct or incorrect, which leaves students to try to figure out what skills or concepts they need to study.

In contrast to traditional assessments, CAT dynamically adjusts question difficulty in real time based on examinees' responses, yielding benefits such as heightened assessment accuracy, testing efficiency, and improved test security [8]. By identifying student proficiency and adaptively selecting questions close to the student's proficiency, CATs continually deliver questions within the student's zone of proximal development [29,30]. The optimal question selection allows for cognitive diagnostic models (detailed below) to disentangle the various contribution to question difficulty, and therefore more accurately measure skill mastery. Additional models can also be added to determine the number of questions that assesses each concept and each content area. Finally, while extensive student data and large question banks are still needed to develop CATs, we have developed CATs that are available online for all courses and institutions, which eliminates the need for individual instructors or institutions to create their own CATs.

Computer Adaptive Testing. CAT assessments typically begin by estimating that a student is at an average ability level. The algorithm selects an item most appropriate for the estimated ability level. Once the student responds to the item, the ability level is adjusted up or down based on the correctness of the student answer. The algorithm then selects a new item that that is most appropriate for the new ability estimate. As the CAT progresses the student ability estimate becomes more stable, and the uncertainty of the ability estimate decreases (Figure 1).

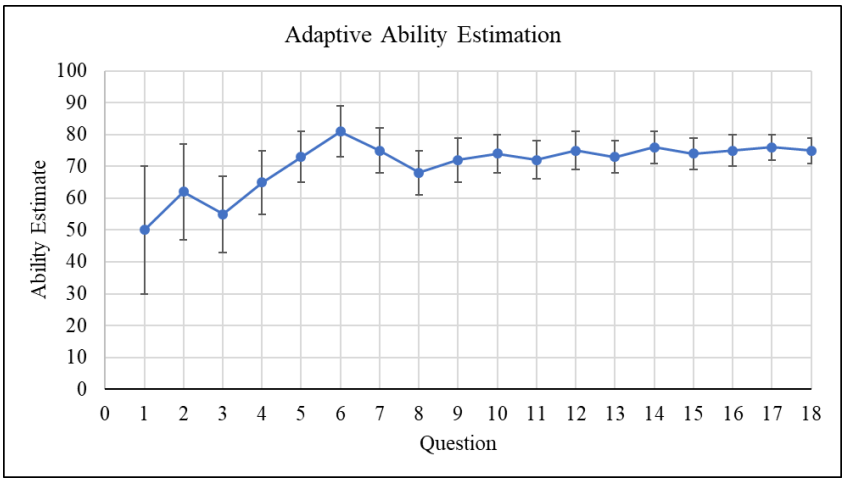


Fig 1. Example of dynamic ability estimation in CAT. Note both the stabilization of the ability estimate and the reduction of uncertainty as indicated by the error bars.

The process for building a CAT platform begins with the calibration of item parameters and the estimation of person parameters, following the principles outlined in various Item Response Theory (IRT) models [31,32]. These models involve one to four parameters that describe characteristics of logistic curves. For instance, three-parameter (3PL) IRT models describe the logistic curve using three item parameters, 'a', 'b', and 'c'. The 'a' parameter represents item discrimination, the 'b' parameter represents item difficulty, and the 'c' parameter represents the guessing parameter, along with person parameters represented by 'θ' for ability[1] as a latent trait.

Parameters for each item are initially calibrated through pretesting where large numbers of students answer each item. This allows the item parameter to be calculated independent of estimation of individual ability levels. Once the characteristics of each item are calculated, the ability levels of test takers can be estimated from their response patterns, usually through maximum likelihood estimation. In other words, the simplest strategy for developing and delivering CAT utilizes an existing item bank that has been given to a large and diverse sample of students. This allows the existing items to be calibrated with IRT models to provide the item difficulty, item discrimination level, and item guessing [1, 32-35].

Once items have been calibrated, a method for selecting items for individual examinees is needed. Various item selection methods, including the fisher information criterion and others, facilitate optimal item selection at each step, aligning the difficulty level with the current ability estimate. In other words, the algorithm tries to find any item from the item bank that has the peak of the information, and this maximization happens in a specific difficulty value which is equal to very close to the ability of the students. This method tries to make the estimated ability close to the item difficulty [3,36,37].

CAT Administration. CAT stands at the forefront of educational assessment, offering a nuanced approach that ensures each student is appropriately challenged, thereby providing a more

[1] We use the term ability consistent with its use in psychometrics to represent the latent construct estimated by CAT, however, we feel that the term “proficiency” better reflect our position that students’ physics knowledge and problem-solving capabilities can evolve over time.

accurate reflection of their knowledge and abilities [38]. CAT integrates a variety of algorithms and strategies, each designed to address specific concerns such as mastery level or overall ability level. This multifaceted approach allows CAT to provide diverse types of immediate feedback, fostering timely reflection and learning. This feedback, crucial for reinforcing conceptual understanding, facilitates the early identification of misconceptions and can be tailored to both item-level performance and overall student ability. The strategies employed in CAT predominantly draw from Item Response Theory (IRT) and are further enhanced by Cognitive Diagnostic Models (CDMs), which focus on assessing students' mastery levels. These comprehensive strategies in CAT effectively manage exam duration concerns, reducing the number of questions required for an accurate assessment of a student [38,39]. Additionally, all questions used with the CAT for this project are assessed for question fairness by calculating differential item functioning between various genders and ethnicities.

Cognitive Diagnostic Models. CDMs offer a detailed view that complements IRT in educational evaluations, providing a nuanced and precise understanding of individual capabilities and skill mastery [40,41]. While IRT targets latent traits to estimate student proficiency, CDMs use latent classes to classify students by their mastery of underlying skills [40,42,43]. Skills cut across content areas and multiple skills may be needed to correctly solve an item. CDMs are classification models that aim to classify a student's skill mastery for predetermined skills identified by content experts (See Table 1 for definitions). The skills require to correctly answer a question are coded dichotomously within matrix called the Q-matrix, which is used by DCMs to estimate mastery. While there are many CDMs, the simplest is the Deterministic Inputs, Noisy "And" gate (DINA) model. DINA is a crucial cognitive diagnostic tool to effectively estimate skill mastery, such as proficiency in applying vectors in physics, providing a detailed understanding of a learner's specific competencies [44]. This model assumes that correctly answering an item in the DINA model hinges on satisfying two conditions. First, examinees must possess all the requisite skills and avoid slipping. Second, examinee that lack any of the necessary skills must correctly guess to achieve a correct answer [43,45].

Learning About STEM Student Outcomes (LASSO) Platform. At the school level, items designed by teachers may face challenges in accuracy, reliability, and calibration due to limited response data. Our project, aiming to optimize the benefits of CAT for schools, addresses these challenges by utilizing the data from the LASSO online platform. LASSO is an online tool designed to assist instructors in evaluating their courses [9,46,47]. The platform conducts online assessments for students, allowing for more efficient use of class time, and it automatically analyzes the collected data. LASSO, aligned with national STEM curriculum and offers a rich test bank of traditional concept inventories (Table 2). Using existing conceptual inventories available on LASSO and the large national database of student responses to these questions. LASSO has developed the MCD, a CD-CAT developed to assess student proficiency, conceptual mastery, and skill mastery of physics students enrolled in introductory mechanics courses.

Results

To facilitate the calibration of items for potential inclusion in our item bank, we conducted a preliminary analysis of student responses using data from the LASSO platform. This analysis was carried out before implementing a CAT based on the Item Response Theory (IRT) Three-

Parameter Logistic (3PL) model. In Table 3, we present a sample of parameters and characteristics of the items obtained after fitting the IRT 3PL model. These items exhibit reasonable values for the difficulty parameter (b), falling within the range of -2 to 2, and the discrimination parameter (a), which ranges from 0 to 2. Additionally, the guessing parameter (c), which is also within reasonable limits. Furthermore, the Root Mean Square Error of Approximation (RMSEA) for each item, all below 0.1, signifies that the IRT model demonstrates an acceptable fit, allowing us to retain the majority of the items.

Table 3 reveals the items that remain in the Force Concept Inventory (FCI) and the Force and Motion Conceptual Evaluation (FCME) after applying the 3PL IRT parameters as filters. These assessments encompass a spectrum of skills pertinent to physics education, where each skill is precisely defined as a distinct component of knowledge or expected student behavior [33,34,38].

Table 2: Sample of Available Instruments on LASSO

Physics	Engineering	Math	Psychological Inventories
• Force Concept Inventory (FCI)	• Electromagnetic Concepts Inventory - Fields (EMCIF)	• Calculus 1 Concept Inventory (C1CI)	• Dweck Mindset Instrument (DMI)
• Force and Motion Conceptual Inventory (FCME)	• Electromagnetic Concepts Inventory - Fields & Waves (EMCIFW)	• Calculus 2 Concept Inventory (C2CI)	• Metacognitive Awareness Inventory (MAI)
• Energy and Momentum Conceptual Survey (EMCS)	• Fluid Mechanics Concept Inventory (FMCI)	• Calculus Concept Assessment (CCA)	• The Perceived Group Inclusion Scale (PGIS)
• Colorado Learning Attitudes about Science Survey (CLASS)	• Heat Transfer Concepts Inventory (HTCI)	• Calculus Concept Inventory (CCI)	• Revised Implicit Theories of Intelligence (Self-Theory) Scale (RITIS)
• Brief Electricity and Magnetism Assessment (BEMA)	• Thermodynamics Concept Inventory (TCI)	• Test of Understanding of Vectors (TUV)	

The Computerized Adaptive Testing (CAT) for each student's report involves evaluating their proficiency level and mastery of various predefined skills, relying on both IRT and Cognitive Diagnostic Modeling (CDM). Therefore, our task extends beyond item preparation and IRT analysis to also include item preparation for the CDM DINA model. To achieve this, we convened instructors and experts in physics education to design the Q-matrix, which represents the binary relationship between items and skills necessary to execute the DINA model. Table 4 outlines the Q-matrix for the FCI and FCME items.

Table 3. Sample results of the IRT model of sample questions from FCI and FCME

Item	Estimated Item Parameters			Item Fit	
	a	b	c	RMSEA.S_X2	p.S_X2
FMCE 1	1.475	-0.73	0.453	0.009	0.001
FMCE 2	2.238	0.456	0.356	0.009	0.003
FMCE 3	1.578	-0.225	0.182	0	0.839
FMCE 47	1.256	0.215	0.201	0.009	0.214
FCI 1	1.475	-0.73	0.453	0.009	0.001
FCI 2	2.238	0.456	0.356	0.009	0.003
FCI 3	1.578	-0.225	0.182	0	0.839
FCI 30	1.625	0.332	0.253	0.009	0.224

These skills are aligned with educational standards, learning objectives, or curriculum goals, and they aim to encompass the diverse array of abilities required for mastery within the educational context. This demonstrates how the skills associated with each item were meticulously defined, ensuring that each skill is grounded in at least one foundational skill [48-49].

Table 4. Q matrix for selected items on the FCI and FCME.

Item	Apply Vectors	Select Appropriate Equations	Algebraic Manipulation	Interpret Graphs
FCI_1	1	1	0	1
FCI_2	1	1	0	1
FCI_3	1	1	1	0
FCI_4	0	1	0	0
FCI_5	1	0	0	0
FCI_6	0	1	0	0
FCI_30	0	1	1	0
FMCE_1	1	1	1	0
FMCE_2	1	1	1	0
FMCE_3	1	1	0	0
FMCE_4	0	1	1	1
FMCE_5	1	1	0	0
FMCE_6	0	1	0	0
FMCE_47	0	1	1	0

While we did not employ the CDM DINA model for CAT item selection, we will provide a report indicating mastery and non-mastery. As we developed the CDM DINA model after the IRT 3PL model, we analyzed the remaining FCI and FCME items using the CDM DINA. In Table 5, we present the parameters specific to the CDM DINA model, including the Slip and guessing parameters. Notably, the guessing parameters, predominantly falling below 0.6, are deemed reasonable. The overall model fit, as assessed by RMSEA2 (which relies on the M2 test rather than

the Chi2 test), indicates that the model is well-suited to capturing the mastery status based on the existing data, as well as for data collected after the CAT administration [44].

In the CAT system within the LASSO framework, the process initiates with the administration of questions of medium difficulty. This methodology is predicated on the assumption that at the commencement of the assessment, all examinees possess a uniform ability level and a moderate level (mostly $\Theta = 0.5$). Such an initial strategy is instrumental in fostering fairness and equitability within the realm of educational assessments [44].

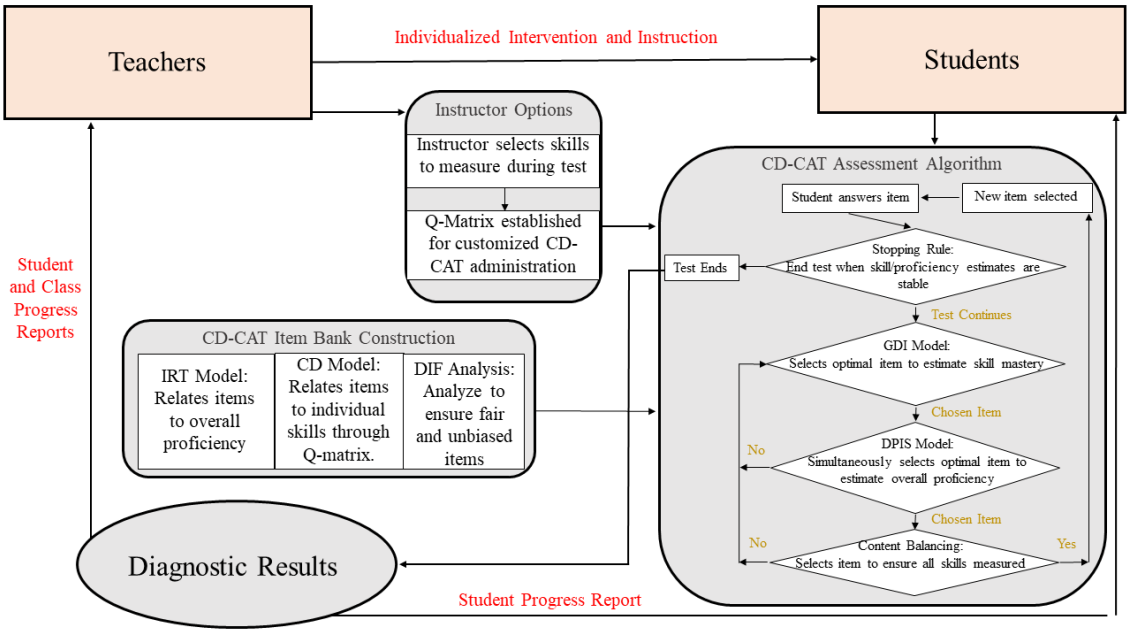
Upon the examinee's response to an initial question, the CAT system engages in a prompt evaluation of the response's accuracy. This step involves updating the ability level—a critical person-level parameter—predicated upon the estimation of likelihood. This immediate assessment is integral to the system's next step - adjusting the difficulty level of forthcoming questions. After this, by the estimated ability, model by using the Fisher information and minimize the differentiate

Table 5. The CDM DINA model parameters for selected items on the FCI and FCME.

Item	guessing	slip	
FCI_1	0.7342	0.0449	RMSEA2
FCI_2	0.4291	0.1921	0.057
FCI_3	0.4612	0.0094	AIC
FCI_4	0.4848	0.071	517766.1
FCI_5	0.2132	0.2367	BIC
FCI_6	0.7059	0.0357	518097.1
FMCE_1	0.2525	0.0998	RMSEA2
FMCE_2	0.1336	0.179	0.078
FMCE_3	0.2501	0.1863	AIC
FMCE_4	0.2197	0.1356	281387.1
FMCE_5	0.2793	0.1736	BIC
FMCE_6	0.0713	0.6805	282243.1

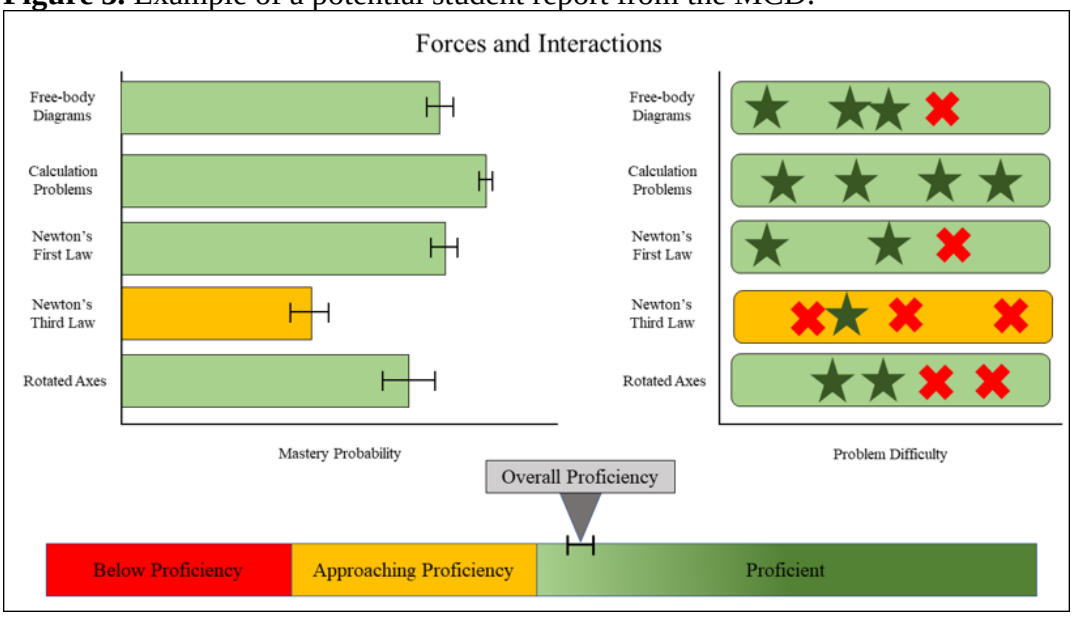
of information matrix and item difficulty. Optimal selection occurs when an item is identified whose difficulty closely approximates the examinee's estimated ability from the preceding item. The reason is that Fisher's information becomes maximum around the point that ability (theta) is equal to the difficulty. However, a predicted issue here is to select the items from a full item bank regardless of the item content. So, in this step, our algorithm tries to check the minimum number of items for each content and if this requirement is met, the algorithm will move on to select from other contents. To address this, the algorithm incorporates a mechanism to ensure a minimum representation of each content category, transitioning to alternate content areas once this criterion is satisfied. This approach emphasizes the importance of adaptively choosing items that are most informative about the examinee's current ability level [44,50,51].

Figure 2. Flowchart of CD-CAT in LASSO



This adaptive process is iterative and dynamic, significantly controlling the precision of the assessment. The continuation of this process is governed by predefined termination criteria, which may include reaching a specified number of questions, exhausting the maximum allotted testing duration, or arriving at a conclusion when the algorithm ascertains that the estimate of the examinee's ability has attained a sufficient level of precision. At the culmination of the CAT, the system generates a final estimate of the examinee's ability. In line with the objective of CD-CAT to ascertain students' skill mastery, post-CAT responses are analyzed through the CDM DINA model to evaluate mastery and non-mastery statuses. To illustrate the operational efficacy and mechanics of the CD-CAT system, Figure 1 illustrates the proficiency estimation over the first 18 questions of a theoretical student taking the MCD. Figure 2 shows the CD-CAT process in LASSO.

Figure 3. Example of a potential student report from the MCD.



Discussion

The deployment of the MCD using a CD-CAT framework and delivered on the LASSO platform marks a significant evolution in educational assessments. CAT's ability to adjust question difficulty in real-time to individual responses enhances both the accuracy and efficiency of evaluations, presenting a promising advancement [1,4,52,53]. Yet, integrating CAT across diverse educational landscapes, particularly in subjects requiring deep conceptual understanding, presents several challenges. These include maintaining the question bank's integrity, ensuring adaptive algorithms' accuracy, and aligning evaluations with curricular standards [1,14]. This shift necessitates a reevaluation of teaching strategies to complement CAT's sophisticated assessment capabilities effectively.

Efforts to broaden CAT's application aim to transform traditional assessment practices, offering a tailored evaluation of student abilities. Preliminary results from CAT implementations have been promising, indicating enhanced student learning outcomes. Recognizing these challenges, the project focuses on large-scale CAT deployment and content balance, aligning with current educational research [1,4,53], thereby enriching our understanding of CAT's impact across different subjects.

While the MCD has been designed and implemented within LASSO and the preliminary analyses look very promising, future work is critical for documenting the efficacy and educational impact of this CD-CAT. We are currently engaging in end user testing to examine how best to deliver information about student proficiency, and concept, content, and skill mastery for both individual students and whole classes. Figure 3 presents an example individual student report showing both proficiency and concept mastery. Skill level mastery could similarly be displayed using bars. Future directions for the MCD and other CD-CATs that will be hosted on LASSO include collecting and analyzing feedback from both students and instructors to examine the impact on student outcomes. In addition, in depth qualitative studies will be needed to examine how instructors can most effectively use the information from the CD-CATs to target interventions to student. This feedback is crucial for refining CAT to meet user needs and improve learning experiences.

Because this platform is widely available for a variety of educational contexts (e.g., undergraduate courses for majors and non-majors, or high school courses) and across many institution types (e.g., HBCUs, HSIs, two-year and four-year institutions), future research will examine the impact of individualized feedback on diverse learners. In addition, future research in customizing potential supplemental instruction that is matched to mastery profiles is essential for large scale implementation. Expanding CD-CAT use across the LASSO platform also requires CD-CAT development for additional subjects (e.g., Chemistry or Math) and for additional grade levels. Future research and development is needed to encompass these subjects and test the versatility and effectiveness of CD-CAT across various educational settings.

References

[1] E. Istiyono, W. S. B. Dwandaru, Y. A. Lede, F. Rahayu, and A. Nadapdap, “Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge,”

International Journal of Instruction, *12*(4), 267–280. 2019. https://doi.org/10.29333/iji.2019.12417a
[2] R. Daphine, P. Sivakumar, and S. Selvakumar, S. "A study on student's Attitude towards online Computer Adaptive Test (CAT) in Physics Education through Observation Schedule," *Journal of Xidian University*, *14*(5), pp.4703-4708. 2020.
[3] H. Wainer, *Computerized adaptive testing: A primer, 2nd ed.* Mahwah, NJ: Erlbaum, 1998. https://psycnet.apa.org/record/2000-03637-000
[4] J. W. Morphew, J. P Mestre, H. A. Kang, H.-H. Chang, and G. Fabry, "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. 2018. https://doi.org/10.1103/PHYSREVPHYSEDUCRES.14.020110/
[5] H.-H. Chang, "Psychometrics behind computerized adaptive testing," *Psychometrika*, 80, 1–20, 2015.
[6] D. J. Weiss, "Improving measurement quality and efficiency with adaptive testing," *Applied Psychological Measurement*, 6, 473–492, 1982.
[7] A. Sahin and D. Ozbasi, "Effects of content balancing and item selection method on ability estimation in computerized adaptive tests," *Eurasian Journal of Educational Research*, 69, 21-36, 2017.
[8] S.-Y. Chen, P.-W. Lei, and W.-H. Liao, "Controlling item exposure and test overlap on the fly in computerized adaptive testing," *British Journal of Mathematical and Statistical Psychology*, 61, 471–492, 2008.
[9] Learning Assistant Alliance, *Learning About STEM Student Outcomes (LASSO)*, 2024. https://learningassistantalliance.org/public/lasso.php
[10] M. L. Loughry, M. W. Ohland, and D. J. Woehr, "Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools," *Journal of Marketing Education*, *36*(1), 5–19. 2014. https://doi.org/10.1177/0273475313499023
[11] M. R. Ab Hamid, W. Sami, and M. H. Mohamad Sidek, "Discriminant Validity Assessment: Use of Fornell & Larcker criterion versus HTMT Criterion," *Journal of Physics: Conference Series*, 890, 012163, 2017. https://doi.org/10.1088/1742-6596/890/1/012163
[12] L. J. Shuman, M. Besterfield-Sacre, and J. McGourty, "The ABET "professional skills" - Can they be taught? Can they be assessed?," *Journal of Engineering Education*, *94*(1), 41–55, 2005. https://doi.org/10.1002/J.2168-9830.2005.TB00828.X
[13] A. Willmott, "Assessment and performance," *Oxford Review of Education*, *4*(1), 51–64, 1978. https://doi.org/10.1080/0305498780040105
[14] P. Zhan, W. Ma, H. Jiao, and S. Ding, "A Sequential Higher Order Latent Structural Model for Hierarchical Attributes in Cognitive Diagnostic Assessments," *Applied Psychological Measurement*, *44*(1), 65, 2020. https://doi.org/10.1177/0146621619832935
[15] D. B. Rivera, C. C. Kuehne, and M. M. Banbury, "Performance-Based Assessment," *Gifted Child Today*, *18*(5), 34–40, 1995. https://doi.org/10.1177/107621759501800511
[16] R. Glaser, N. Chudowsky, and J. W. Pellegrino, *Knowing what students know: The science and design of educational assessment*. Washington D.C.: National Academies Press, 2001.
[17] M. K. Demir and M. Y. Eryaman, "A qualitative evaluation of instructors' exam questions at a primary education department in terms of certain variables," *Educational Policy Analysis and Strategic Research*, 7(1), 52-63, 2012.

[18] C. D. Wright, A. L. Huang, K. M. Cooper, and S. E. Brownell. "Exploring differences in decisions about exams among instructors of the same introductory biology course," *International Journal for the Scholarship of Teaching and Learning* 12(2), 14, 2018.
[19] J. C. McNeil and M. W. Ohland, "Engineering faculty perspectives on the nature of quality teaching," Quality Approaches in Higher Education, 6(2), 20-30, 2015.
[20] E. Broemer, and G. Recktenwald, Cheating and Chegg: A retrospective. In *2021 ASEE Virtual Annual Conference*, Paper #34650, 2021, July.
[21] A. Chirumamilla, G. Sindre, and A. Nguyen-Duc, "Cheating in e-exams and paper exams: the perceptions of engineering students and teachers in Norway," *Assessment & Evaluation in Higher Education*, *45*(7), 940-957. 2020.
[22] B. E. Whitley. "Factors associated with cheating among college students: A review," *Research in Higher Education*, 39, 235–274, 1998.
[23] M. P. Watters, P. J. Robertson, and R. K. Clark. "Student perceptions of cheating in online business courses," *Journal of Instructional Pedagogies*, 6, 2010.
[24] M. P. Escudier, T. J. Newton, M. J. Cox, P. A. Reynolds, and E. W. Odell, "University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a 'high-stakes' examination," *Journal of Computer Assisted Learning*, *27*(5), 440-447, 2011.
[25] C. A. Emeka, C. Zilles, M. West, G. L. Herman, and T. Bretl, "Second-Chance Testing as a Means of Reducing Students' Test Anxiety and Improving Outcomes," In *2023 ASEE Annual Conference & Exposition*, 2023, June.
[26] F. Moosvi, D. Eddelbuettel, C. Zilles, S. A. Wolfman, F. Fund, L. K. Alford, and J. Schroeder, "Creating Algorithmically Generated Questions Using a Modern, Open-sourced, Online Platform: PrairieLearn." In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, 1177-1177, 2022, March.
[27] M. Fowler, D. H. Smith IV, C. Emeka, M. West, and C. Zilles, "Are we fair? quantifying score impacts of computer science exams with randomized question pools," In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, 647-653, 2022, February.
[28] J. W. Morphew, M. Silva, G. L. Hermann, and M. West, "Frequent mastery testing with second chance exams leads to enhanced student learning in undergraduate engineering," *Applied Cognitive Psychology*, *34*, 168-181, 2019. DOI: 10.1002/acp.3605
[29] N. D. Fila, T. M. Fernandez, S. Purzer, and A. S. Bohlin, "Innovation and the zone of proximal development in engineering education," In *2016 ASEE Annual Conference & Exposition*, 2016, June.
[30] B. Eun, "The zone of proximal development as an overarching concept: A framework for synthesizing Vygotsky's theories," *Educational Philosophy and Theory*, *51*(1), 18-30, 2019.
[31] L. Laatsch and J. Choca, "Cluster-branching methodology for adaptive testing and the development of the adaptive category test," *Psychological Assessment, 6*(4), 345–351, 1994. https://doi.org/10.1037/1040-3590.6.4.345
[32] C. Hasse, "Postphenomenology: Learning cultural perception in science," *Human Studies*, *31*(1), 43–61, 2008. https://doi.org/10.1007/S10746-007-9075-4/METRICS
[33] E. Istiyono, W. Sunu, B. Dwandaru, and R. Faizah, "Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT," *REID: Research and Evaluation in Education*, *4*(2), 144–154, 2018. https://doi.org/10.21831/REID.V4I2.22218

[34] B. Ozdemir and S. Gelbal, “Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study,” *Education and Information Technologies*, *27*(5), 6273–6294, 2022. https://doi.org/10.1007/S10639-021-10853-0/FIGURES/4
[35] S. Saarinen, E. Cater, and M. L. Littman, “Applying prerequisite structure inference to adaptive testing,” *ACM International Conference Proceeding Series*, 422–427, 2020. https://doi.org/10.1145/3375462.3375541
[36] B. Keskin and M. Gunay, "A survey on computerized adaptive testing," *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Elazig, Turkey, 1-6, 2021. https://doi.org/10.1109/ASYU52992.2021.9598952
[37] L. H. Thamsborg, M. A. Petersen, N. K. Aaronson, W. C. Chie, A. Costantini, B. Holzner, I. M. V. de Leeuw, T. Young, and M. Groenvold, “Development of a lack of appetite item bank for computer-adaptive testing (CAT),” *Supportive Care in Cancer*, *23*(6), 1541–1548. 2015. https://doi.org/10.1007/S00520-014-2498-3
[38] W. J. Van Der Linden “Conceptual Issues in Response-Time Modeling,” *Journal of Educational Measurement*, *46*(3), 247–272, 2009. https://doi.org/10.1111/J.1745-3984.2009.00080.
[39] M. A. van der Kooij, “The impact of chronic stress on energy metabolism,” *Molecular and Cellular Neurosciences*, 107, 2020. https://doi.org/10.1016/J.MCN.2020.103525
[40] J. Liu, W. Tang, X. He, B. Yang, and S. Wang, “Research on DINA Model in Online Education,” in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, A. Mobasheri, Ed. Cham: Springer International Publishing, 2020, pp. 279–291. https://doi.org/10.1007/978-3-030-63955-6_24
[41] E. Thompson, A. Luxton-Reilly, J. L. Whalley, M. Hu, and P. Robbins, P. “Bloom’s Taxonomy for CS Assessment,” in *Proceedings of the tenth conference on Australasian computing education-Volume 78*, 2008, pp. 155-161.
[42] T. O. Başokçu, T. Öğretmen, and H. Kelecioğlu, “Model Data Fit Comparison between DINA and G-DINA in Cognitive Diagnostic Models,” *Education Journal*, *2*(6), 256, 2013. https://doi.org/10.11648/J.EDU.20130206.18
[43] J. de la Torre, “The Generalized DINA Model Framework,” *Psychometrika*, *76*(2), 179–199, 2011. https://doi.org/10.1007/S11336-011-9207-7
[44] H. Ravand, “Cognitive Diagnostic Modeling Using R,” *Practical Assessment, Research & Evaluation*, *20*(1), 11, 2015.
[45] J. de la Torre, “DINA Model and Parameter Estimation: A Didactic,” *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130, 2009. https://doi.org/10.3102/1076998607309474
[46] B. Van Dusen, “LASSO: A new tool to support instructors and researchers,” *American Physics Society Forum on Education*, 2018, Fall.
[47] J. M. Nissen, I. Her Many Horses, B. Van Dusen, M. Jariwala, and E. Close, “Providing context for identifying effective introductory mechanics courses,” *The Physics Teacher*, 60, 179–182, 2022.
[48] Chandio, M. T., Pandhiani, S. M., & Iqbal, R. (2016). Bloom’s Taxonomy: Improving Assessment and Teaching-Learning Process. *Journal of Education and Educational Development*, *3*(2), 203–221.

[49] L. C. Sanchez and B. L. Maribao DIT, “Computer adaptive testing using iterative algorithm,” *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3870–3876, 2020. https://doi.org/10.30534/ijatcse/2020/206932020

[50] P. Gilavert and V. Freire, V. “Computerized adaptive testing: A unified approach under Markov Decision Process,” in *International Conference on Computational Science and Its Applications*, Cham: Springer International Publishing, 2022, pp. 591-602. https://doi.org/10.1007/978-3-031-10522-7_40

[52] J. Pacheco-Ortiz, L. Rodríguez-Mazahua, J. Mejía-Miranda, I. Machorro-Cano, and U. Juárez-Martínez, “Towards association rule-based item selection strategy in computerized adaptive testing,” *Studies in Computational Intelligence*, *966*, 27–54, 2021. https://doi.org/10.1007/978-3-030-71115-3_2

[53] A. Z. Abidin, E. Istiyono, N. Fadilah, W. Sunu, and B. Dwandaru, “A computerized adaptive test for measuring the physics critical thinking skills in high school students,” *International Journal of Evaluation and Research in Education*, *8*(3), 376–383, 2019. https://doi.org/10.11591/ijere.v8i3.19642

[54] U. C. Müller, T. Huelmann, M. Haustermann, F. Hamann, E. Bender, and D. Sitzmann, “First results of computerized adaptive testing for an online physics test,” in *Towards a New Future in Engineering Education, New Scenarios That European Alliances of Tech Universities Open Up*, Universitat Politècnica de Catalunya, 2022, pp. 1377–1387. https://doi.org/10.5821/CONFERENCE-9788412322262.1273