Mechanics Cognitive Diagnostic: Mathematics skills tested in introductory physics courses

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

Jayson M. Nissen
Nissen Education and Research Design, Monterey, CA, 93940, USA

Xiuxiu Tang, Yuxiao Zhang, and Hua Hua Chang
College of Education, Purdue University, West Lafayette, IN, 47907, USA

Jason W. Morphew
School of Engineering Education, Purdue University, West Lafayette, IN, 47907, USA

Physics instructors and education researchers use research-based assessments (RBAs) to evaluate students' preparation for physics courses. This preparation can cover a wide range of constructs, including mathematics and physics content. Using separate mathematics and physics RBAs consumes course time. We are developing a mechanics cognitive diagnostic (MCD) as an online test using both computerized adaptive testing and cognitive diagnostic models. This design allows the MCD to assess mathematics and physics content knowledge within a single assessment. Our work used an evidence-centered design framework to inform the extent to which our models of skills students develop in physics courses fit the data from three mathematics RBAs. Our dataset came from the LASSO platform and includes 3,491 responses from the Calculus Concept Assessment, Calculus Concept Inventory, and Pre-calculus Concept Assessment. Our model included five skills: apply vectors, conceptual relationships, algebra, visualizations, and calculus. The "deterministic inputs, noisy 'and' gate" (DINA) analyses demonstrated a good fit for the five skills. The classification accuracies for the skills were satisfactory. Including items of the three mathematics RBAs in the MCD's item bank will provide a flexible assessment of these skills across mathematics and physics content areas that can adapt to instructors' needs.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Le

I. INTRODUCTION

Instructors and researchers often use research-based assessments (RBAs) to evaluate students' performance and the effectiveness of different pedagogies in physics education research [1–3]. For example, the Force Concept Inventory (FCI) [4] assesses students' understanding of force concepts, and the Calculus Concept Inventory (CCI) [5] assesses calculus knowledge; both are important in introductory physics courses [6]. By assessing students' knowledge, instructors can adjust their teaching approaches to better fit their student's needs, and researchers can identify effective pedagogical practices [7] and inequities in the course outcomes [8–10].

Mathematics ability is foundational for learning in physics [11]. Some physics instructors use mathematics RBAs to assess the student's mathematical readiness level and comprehension of mathematical concepts [1]. For instance, PhysPort (a website with resources for teaching physics) provides nine mathematics RBAs appropriate for physics courses [12]. Instructors can use these RBAs to better support students in acquiring the mathematics knowledge they need to succeed in their physics courses.

Using mathematics RBAs poses physics instructors and researchers with the dilemma of choosing what to measure: specific physics content knowledge, mathematics content knowledge, or affective measures like self-efficacy. All RBAs on PhysPort, for example, use fixed-length designs that require using the whole assessment or risking undermining the validity of arguments for the assessment. This design and length also limit using these assessments in a time-effective manner, such as for weekly quizzes. Reducing the test length while maintaining test validity requires extensive work. Computerized adaptive testing (CAT) can address this issue by optimizing test length [13] to accurately measure each student's proficiency level [14].

We are developing the mechanics cognitive diagnostic (MCD) as an online adaptive RBA for introductory physics courses using both CAT and cognitive diagnostic (CD) models [15]. The CAT selects the next item based on a student's previous responses to obtain the most information possible [16]. The CD model measures student mastery of several skills that students need to succeed in physics. The efficiency of the CAT allows the CD models to assess many skills. The CD model uses a Q-matrix to link test items to each skill to generate a profile of an individual's skill mastery [17, 18]. The MCD, thus, enables instructors or researchers to accurately and efficiently assess both overall student proficiency and specific skill mastery.

The MCD assesses four skills for the physics content areas (i.e., apply vectors, conceptual relationships, algebra, and visualizations)[15], which comprise the student model (see Sec. II). The MCD item bank largely draws physics items from the Force Concept Inventory [4], Force and Motion Conceptual Evaluation [19], and Energy and Momentum Conceptual Survey[20]. To extend the validity arguments for the MCD's student model and to identify mathematics items for inclusion in the MCD, we asked the following research question for three mathematics RBAs: the Calculus Concept Assessment (CCA) [21], the Calculus Concept Inventory (CCI) [5], and the Pre-calculus Concept Assessment (PCA) [22].

- *What skills from our student models, if any, did the three mathematics RBAs measure?*

To support readers' interpretation of our research, Table I includes a selection of terms and their definitions [15].

II. THEORETICAL FRAMEWORK

We used evidence-centered design (ECD) [30] to inform our development of the MCD, concentrating on collecting evidence to support claims about a student's skills [31]. The ECD consists of five models (see Fig. 1) to provide a framework for designing and developing RBAs. To address our research question, we focused on the student and evidence models from the ECD to assess skills in the mathematics content areas. The student models aim to determine the variables (i.e., skills and content areas) related to performance that our assessment seeks to measure, ensuring that the assessment aligns with our intended goals. The evidence models, which consist of evidence rules and the measurement model as sub-models, establish the criteria for evaluating an RBA. Although not the focus of this investigation, we will provide an overview of parts (3)-(5) of the ECD as they offer relevant context for this work. The task model is centered around multiple-choice questions, each with a definitive right or wrong answer. The assembly model integrates the student, evidence, and task models to estimate the overall students' proficiency and their skill mastery profile. The delivery system model uses the LASSO platform [32, 33] to administer the assessment online and address test security and timing.

III. MATERIALS AND METHODS

The LASSO platform provided the dataset for this research purpose [32]. The dataset consists of 3,493 first-year college student responses across three post-course assessments: the CCA (17 items - 1,292 students), CCI (22 items - 940 students), and PCA (25 items - 1,259 students). Four institutions used the CCA in 53 courses, seven used the CCI in 113 courses, and eight used the PCA in 40 courses. Mathematics courses used all three RBAs while data also came from four physics courses that used the CCA or CCI. The analysis excluded students who took less than five minutes to complete the assessment or did not answer all of the items. In cases where students completed the same assessment multiple times, we only used their first response recent post test.

We used an iterative, mixed-methods approach to analyze the extent to which the student models of skills fit the three mathematics RBAs (CCA, CCI, and PCA). We started with the four skills developed for the physics content: apply vectors, conceptual relationships, algebra, visualizations. We de-

TABLE I. 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 [14, 16, 23].
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. Each row represents a test item, and each column represents a specific skill. An entry of 1 in the matrix indicates that a particular skill is assessed by the corresponding item, while a 0 indicates that the skill is not assessed.
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 [24, 25].
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. [26].
Deterministic inputs, 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 [25, 27–29].
Evidence-centered design (ECD) model - A framework for developing educational assessments based on establishing logical, evidence-based arguments [30].

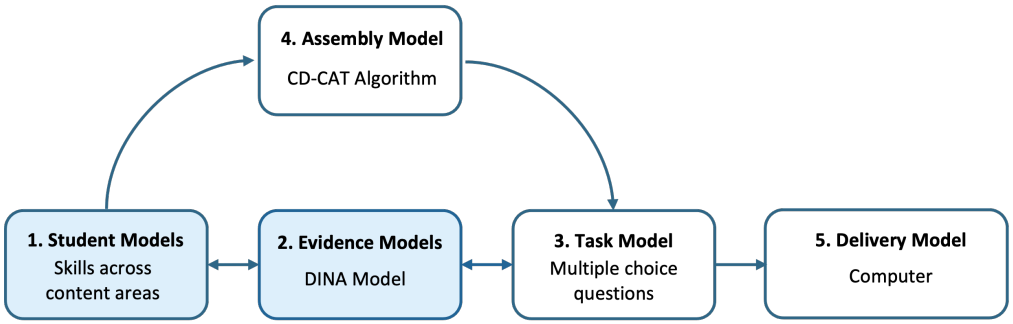

FIG. 1. An evidence-centered design (ECD) model for the creation of mechanics cognitive diagnostic (MCD) [15].

veloped these skills based on learning objectives from physics courses using standards-based grading, e.g., [34], and tested them with data from three physics RBAs [15]. We added the calculus skill, which was not assessed by the physics RBAs, to ensure the student model covered all of the items on the mathematics RBAs. In total, we coded the items for the five skills in our student models [15]: 1) apply vectors, 2) conceptual relationships, 3) algebra, 4) visualizations, and 5) calculus (see Table II). We then empirically tested the fit of these codes using the "deterministic inputs, noisy 'and' gate" (DINA) model (see Table I). The input of the DINA model is a Q-matrix, which specifies the relationship between test items and required skills using binary entries. Each row represents a test item, and each column corresponds to a skill, indicating whether a skill is needed for an item (see Table III). One of the outputs from the DINA model provides a suggested Q-matrix for each RBA to refine our qualitative coding. The empirical results for each RBA suggested changes to the initial qualitative coding. The coders then reviewed these suggested changes and agreed on accepting or rejecting each suggestion. The process concluded with a final DINA analysis of the updated codes.

Our coding team consisted of three researchers with physics backgrounds. Coders independently coded items to create Q-matrices. For the initial coding, we compared our codings to reach a consensus on each item. After running the DINA analysis, we reviewed the suggested Q-matrices of the DINA model, and came to a consensus on accepting or rejecting each proposed change. The final consensus coding of the three RBAs provided inputs into the final DINA analysis presented herein.

We used the G-DINA package in R to analyze our data [35]. Two fit indices were calculated to assess the DINA models fit: the root mean square error of approximation based on the M2 statistic (RMSEA2) and the standardized root mean square residual (SRMR). Lower values of RMSEA2 and SRMR indicate a better fit. Models with RMSEA2 values lower than 0.06 indicate a good fit and up to 0.08 represent an acceptable fit [36]. Models with SRMSR values below 0.05 indicate a good fit and up to 0.08 represent an acceptable fit [37, 38]. The DINA model produces classification accuracy scores that quantify the accuracy of the model's estimation of students' skill mastery. Classification accuracies range from 0-1, with values greater than or equal to 0.9 considered high [39, 40] and values greater than 0.8 are acceptable [41].

TABLE II. Definition of the skills present in the three RBAs.

Skills	Definition
Apply Vectors	Item requires manipulating vectors in more than one dimension or has a change in sign for a 1-D vector quantity.
Conceptual Relationships	Item requires students to identify a relationship between variables and/or the situations in which those relationships apply.
Algebra	Item requires students to reorganize one or more equations. This goes beyond recognizing the standard forms of equations.
Visualizations	Item requires extracting information from or creating formal visualizations such as xy plots, bar plots, or line graphs.
Calculus	Item requires applying limits, derivatives, or integrals (i.e., rates of change.)

TABLE III. The table provides the Q-matrix for the first five CCI items. A 1 indicates the item measured the skill, and 0 is otherwise.

Item	Apply Vectors	Conceptual Relationships	Algebra	Visualizations	Calculus
1	0	0	1	0	0
2	0	1	1	0	0
3	0	1	0	0	1
4	1	0	0	0	0
5	0	0	0	0	1
6	1	0	0	0	0

TABLE IV. Q-matrix modifications and adoption rates.

	Total Items	Possible Changes	Suggested Changes	Adopted Changes	Adoption Rate	Change Rate
CCA	17	85	3	1	33%	1.2%
CCI	22	110	15	4	27%	3.6%
PCA	25	100	15	1	7%	1.0%
Overall	64	295	33	6	18%	2.0%

TABLE V. Model fit by assessment

	CCA	CCI	PCA	Cutoff
RMSEA2	0.026	0.008	0.034	0.050
SRMSR	0.034	0.033	0.044	0.050

IV. FINDINGS

Two results from our analyses support the student models of the five skills fitting the data well. First, across the three RBAs, the initial DINA model suggested 33 out of 295 possible changes (see Table IV). The coding team adopted six of those suggestions, giving an overall change rate of 2.0%. Second, the final DINA models for each of the three RBAs demonstrated good fit with RMSEA2 less than 0.04 and SRMSR less than 0.05 for all three mathematics assessments (see Table V).

TABLE VI. The distribution of items across the number of skills.

	Total items	N(%)		
		1	2	3
CCA	17	4 (23%)	10 (59%)	3 (18%)
CCI	22	15 (68%)	7 (32%)	0 (0%)
PCA	25	15 (60%)	8 (32%)	2 (8%)
Overall	64	34 (53%)	25 (39%)	5 (8%)

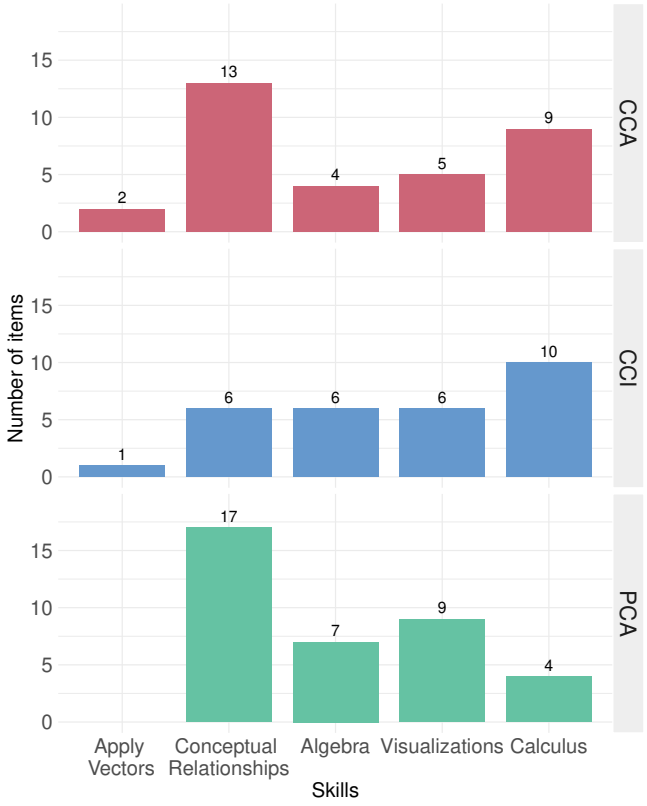

FIG. 2. The distribution of items across skills and assessments.

In terms of which skills the three RBAs assessed, the aggregate number of items assessing conceptual relationships, algebra, visualizations, and calculus skills ranged from 17 to 36 items (see Fig. 2). Only three items in total assessed the apply vectors skill. The sum of the item counts for each assessment in Fig. 2 exceeds the total number of items on each RBA, as shown in Table IV because some items assessed multiple skills. Slightly more than half of the items only assessed one skill (see Table VI) with many assessing two skills and only a few assessing three skills.

The CCA and the CCI assessed all five skills, while the PCA only assessed four of the skills (see Fig. 2). Thus, the DINA model created 14 classification accuracies across the three RBAs that are shown in Table VII. Three of the 14 classification accuracies were below 0.8. The lowest classification accuracies were for the apply vectors skill on the CCA (0.64) and CCI (0.58), which only had two and one

TABLE VII. Skill classification accuracy by assessment.

	Apply Vectors	Conceptual Relationships	Algebra	Visualizations	Calculus
CCA	0.64	0.87	0.86	0.70	0.88
CCI	0.58	0.94	0.94	0.92	0.94
PCA	-	0.92	0.90	0.92	0.84

item, respectively. These results indicate that the MCD will need additional items to assess the apply vectors skill. Seven of the classification accuracies were good (greater than or equal to 0.9) and four classification accuracies were acceptable (greater than 0.8 but less than 0.9). Because the MCD will draw on items from all three of these RBAs, as well as additional future items, to assess these skills, we expect the combined item bank will support acceptable classification accuracies for the MCD.

V. DISCUSSION AND CONCLUSION

The three RBAs contain finer-grained information about students than just their overall performance. Our analysis found that the five-skills student model fit the data well for all three mathematics RBAs. In addition, the DINA model showed high to acceptable classification accuracies for four of the five skills: conceptual relationships, algebra, visualizations, and calculus. The apply vectors skill only had three items across the three RBAs which resulted in the observed low classification accuracy for that skill. We aim to ensure high classification accuracy by including a minimum of 10 items for each combination of content area and skill. This strategy will also make sure that we have enough items to assess student proficiency. Instructors can use this fine-grained, formative assessment to adjust their teaching to meet each student's needs and abilities.

If given student skill mastery profiles, instructors could adjust their instructional strategies in different ways. For instance, consider a student who has mastered algebra but has not mastered vectors skills. Instructors can use this information to assign students instructional interventions that focus on developing the student's ability to apply vectors in physics contexts. Instructors could also create groups of students with complementary skill sets for collaborative assignments or labs. This strategy would support students in learning from each other's strengths while receiving support in their weaker content areas.

Many introductory physics courses expect students to have fluency in using vectors as vectors are foundational to many physics concepts [11]. However, few items on the three mathematics RBAs assessed the apply vectors skill. Physics RBAs, such as the Test of Understanding of Vectors [42] and the Vector Evaluation Test [43], cover vectors in depth. We are developing the MCD by combining mathematics and physics RBAs to provide a large item bank to classify physics students' mastery of all skills.

The findings in this analysis may be limited in their application to the MCD because the MCD focuses on physics courses and content, and these instruments and the analyzed data largely came from mathematics courses. The analyses in this paper also do not reflect how the MCD will operate. Instead, they provide validity evidence for the student models for the MCD and to guide the development of the item bank for the MCD. Subsequent analyses of the MCD will investigate the extent to which the MCD measures the proposed student models of skills.

The CCI and PCA were able to measure multiple skills with high classification accuracies, four and three respectively. Physics instructors, however, may want to assess their students across several different constructs: mathematics and physics conceptual knowledge and affective traits such as self-efficacy or attitudes about experiments. The more time students spend taking these assessments the less time they have for other learning activities and the lower the quality of their responses. We are developing the mechanics cognitive diagnostic (MCD) in the online LASSO platform to address this issue. The MCD minimizes test length while still maintaining measurement accuracy by selecting items for students to complete that provide the most information. Instructors can use the MCD to test these skills across content areas in physics and mathematics. Our purpose is to create a larger MCD item bank, including three physics RBAs, three mathematics RBAs, and others to address the skill gaps, within the context of teaching introductory physics.

The adaptive nature of the MCD allows instructors to use it as a pre/post test at the beginning and end of a course or as a weekly formative assessment. Instructors choose which skills and content areas to assess and when to assess them. This flexibility provides timely feedback on individual student's learning. Researchers can also use this longitudinal data across skills and content areas to model student learning trajectories and identify effective learning activities for each location on that trajectory.

[1] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource letter rbai-2: Research-based assessment instruments: Beyond physics topics, American Journal of Physics **87**, 350 (2019).

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

[3] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning in-

creases student performance in science, engineering, and mathematics, Proceedings of the National Academy of Sciences **111**, 8410 (2014).
[4] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The physics teacher **30**, 141 (1992).
[5] J. Epstein, Development and validation of the calculus concept inventory, in *Proceedings of the ninth international conference on mathematics education in a global community*, Vol. 9 (Citeseer, 2007) pp. 165–170.
[6] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: research-based assessment instruments in physics and astronomy, American Journal of Physics **85**, 245 (2017).
[7] J. M. Nissen, I. H. Many Horses, B. V. Dusen, M. Jariwala, and E. Close, Providing context for identifying effective introductory mechanics courses, The Physics Teacher **60**, 179 (2022).
[8] B. Van Dusen and J. Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, Journal of Research in Science Teaching **57**, 33 (2020).
[9] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Physical Review Special Topics-Physics Education Research **5**, 010101 (2009).
[10] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, Journal of Research in Science Teaching **49**, 1096 (2012).
[11] S. W. Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics inventory of quantitative literacy: A tool for assessing mathematical reasoning in introductory physics, Physical Review Physics Education Research **17**, 020129 (2021).
[12] Physport: Browse assessments (n.d.).
[13] J.-i. Yasuda, N. Mae, M. M. Hull, and M.-a. Taniguchi, Optimizing the length of computerized adaptive testing for the force concept inventory, Physical review physics education research **17**, 010115 (2021).
[14] 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**, 020110 (2018).
[15] V. Le, J. M. Nissen, X. Tang, Y. Zhang, A. Mehrabi, J. W. Morphew, H. H. Chang, and B. Van Dusen, Applying cognitive diagnostic models to mechanics concept inventories, arXiv preprint arXiv:2404.00009 (2024).
[16] H.-H. Chang, Psychometrics behind computerized adaptive testing, Psychometrika **80**, 1 (2015).
[17] Y. Chen, J. Liu, G. Xu, and Z. Ying, Statistical analysis of Q-matrix based diagnostic classification models, Journal of the American Statistical Association **110**, 850 (2015).
[18] Y. Cui, M. J. Gierl, and H.-H. Chang, Estimating classification consistency and accuracy for cognitive diagnostic assessment, Journal of Educational Measurement **49**, 19 (2012).
[19] R. K. Thornton and D. R. Sokoloff, Assessing student learning of newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, american Journal of Physics **66**, 338 (1998).
[20] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, American Journal of Physics **71**, 607 (2003).
[21] Calculus concept assessment (n.d.).
[22] M. Carlson, M. Oehrtman, and N. Engelke, The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings, Cognition and Instruction **28**, 113 (2010).
[23] D. J. Weiss, Improving measurement quality and efficiency with adaptive testing, Applied psychological measurement **6**, 473 (1982).
[24] H. Ravand and A. Robitzsch, Cognitive diagnostic modeling using R, Practical Assessment, Research, and Evaluation **20**, 11 (2015).
[25] J. De La Torre and N. Minchen, Cognitively diagnostic assessments and the cognitive diagnosis model framework, Psicología Educativa **20**, 89 (2014).
[26] W. Wang, L. Song, P. Chen, Y. Meng, and S. Ding, Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment, Journal of Educational Measurement **52**, 457 (2015).
[27] E. Haertel, An application of latent class models to assessment data, Applied Psychological Measurement **8**, 333 (1984).
[28] B. W. Junker and K. Sijtsma, Cognitive assessment models with few assumptions, and connections with nonparametric item response theory, Applied Psychological Measurement **25**, 258 (2001).
[29] J. de la Torre, DINA model and parameter estimation: A didactic, Journal of Educational and Behavioral Statistics **34**, 115 (2009).
[30] R. J. Mislevy, R. G. Almond, and J. F. Lukas, A brief introduction to evidence-centered design, ETS Research Report Series **2003**, i (2003).
[31] B. Pollard, R. Hobbs, R. Henderson, M. D. Caballero, and H. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, Physical Review Physics Education Research **17**, 010133 (2021).
[32] Learning about stem student outcomes (LASSO) (2023).
[33] B. Van Dusen, M. Shultz, J. M. Nissen, B. R. Wilcox, N. Holmes, M. Jariwala, E. W. Close, H. Lewandowski, and S. Pollock, Online administration of research-based assessments, American Journal of Physics **89**, 7 (2021).
[34] I. D. Beatty, Standards-based grading in introductory university physics, Journal of the Scholarship of Teaching and Learning , 1 (2013).
[35] W. Ma and J. de la Torre, GDINA: An R package for cognitive diagnosis modeling, Journal of Statistical Software **93**, 1 (2020).
[36] D. Hooper, J. Coughlan, and M. Mullen, Evaluating model fit: a synthesis of the structural equation modelling literature, in *7th European Conference on research methodology for business and management studies*, Vol. 2008 (2008) pp. 195–200.
[37] L.-t. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural equation modeling: a multidisciplinary journal **6**, 1 (1999).
[38] S. T. Jang, The implications of intersectionality on southeast asian female students' educational outcomes in the united states: A critical quantitative intersectionality analysis, American Educational Research Journal **55**, 1268 (2018).
[39] Z. Tan, J. De la Torre, W. Ma, D. Huh, M. E. Larimer, and E.-Y. Mun, A tutorial on cognitive diagnosis modeling for characterizing mental health symptom profiles using existing item responses, Prevention Science **24**, 480 (2023).
[40] Q. Liang, J. de la Torre, M. E. Larimer, and E.-Y. Mun, Mental health symptom profiles over time: A three-step latent tran-

sition cognitive diagnosis modeling analysis with covariates, Mendeley Data (2023).

[41] J. Paulsen, D. Svetina, Y. Feng, and M. Valdivia, Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models, Applied Psychological Measurement **44**, 267 (2020).

[42] P. Barniol and G. Zavala, Test of understanding of vectors: A reliable multiple-choice vector concept test, Physical Review Special Topics-Physics Education Research **10**, 010121 (2014).

[43] R. K. Thornton, Measuring and improving student mathematical skills for modeling, in *Proceedings GIREP Conference* (2006).