Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course

Jason W. Morphew
Department of Educational Psychology, College of Education, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820, USA

Jose P. Mestre*
Department of Physics, College of Engineering, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820, USA

Hyeon-Ah Kang
Department of Educational Psychology, University of Texas at Austin, Austin, Texas 78712, USA

Hua-Hua Chang
Educational Psychology & Research Methodology, College of Education, Purdue University, West Lafayette, Indiana 47907, USA

Gregory Fabry
West Monroe Partners, 222 W Adams St, Chicago, Illinois 60606, USA

(Received 9 November 2017; published 20 September 2018)

Prior research has established that students often underprepare for midterm examinations yet remain overconfident in their proficiency. Research concerning the testing effect has demonstrated that utilizing testing as a study strategy leads to higher performance and more accurate confidence compared to more common study strategies such as rereading or reviewing homework problems. We report on three experiments that explore the viability of using computer adaptive testing (CAT) for assessing students' physics proficiency, for preparing students for midterm exams by diagnosing their weaknesses, and for predicting scores in midterm exams in an introductory calculus-based mechanics course for science and engineering majors. The first two experiments evaluated the reliability and validity of the CAT algorithm. In addition, we investigated the ability of the CAT test to predict performance on the midterm exam. The third experiment explored whether completing two CAT tests in the days before a midterm exam would facilitate performance on the midterm exam. Scores on the CAT tests and the midterm exams were significantly correlated and, on average, were not statistically different from each other. This provides evidence for moderate parallel-forms reliability and criterion-related validity of the CAT algorithm. In addition, when used as a diagnostic tool, CAT showed promise in helping students perform better on midterm exams. Finally, we found that the CAT tests predicted the average performance on the midterm exams reasonably well, however, the CAT tests were not as accurate as desired at predicting the performance of individual students. While CAT shows promise for practice testing, more research is needed to refine testing algorithms to increase reliability before implementing CAT for summative evaluations. In light of these findings, we believe that more research is needed comparing CAT to traditional paper-and-pencil practice tests in order to determine whether the effort needed to create a CAT system is worthwhile.

DOI: 10.1103/PhysRevPhysEducRes.14.020110

*Corresponding author.
mestre@illinois.edu

I. INTRODUCTION

Computerized adaptive testing (CAT) is a method of administering tests that has become popular in many high-stakes educational testing programs. CAT differs profoundly from traditional paper-and-pencil (P&P) tests. CAT is a dynamic testing procedure where items are selected and administered according to continuously

updating estimates of each examinee's proficiency level. The term proficiency is used throughout this article to indicate an individual's current physics knowledge and problem-solving capabilities as measured using conceptual and calculation-based physics problems. By defining proficiency in this way, we interpret students' proficiency estimates similar to the meaning of student scores on traditional P&P exams.[1]

In contrast, P&P tests are static tests where all examinees are given an identical set of items. A major potential benefit of CAT is that it may be able to provide more efficient estimates of the examinee's proficiency level with fewer items than that required in P&P tests, thereby allowing significant savings in time for test administration compared with P&P tests [1,2]. Benefits of CAT also include easier and faster data analysis, and immediate score reporting [3]. Examples of large-scale, national CAT exams include the Graduate Record Exam (GRE), the National Council of State Boards of Nursing (NCLEX), the National Assessment of Educational Progress (NAEP), and the Armed Services Vocational Aptitude Battery (ASVAB).

To date, the possible benefits of CAT have not been explored in the context of large introductory STEM university courses. For example, CAT could be used as a diagnostic formative assessment for students interested in finding out what score they would receive on a midterm exam prior to taking it. As a diagnostic tool, CAT can provide students with a realistic assessment of their current level of competence by reporting a predicted score. In addition, CAT can also provide students with diagnostic information on the topics or types of problems in which they are weak. This feedback could allow students to focus on those topics and spend less time studying topics that they know well. This type of diagnostic feedback is very important because research consistently shows that students, especially less prepared students, come to course exams overconfident and underprepared, a common phenomenon known as the Dunning-Kruger effect [4–6]. CAT could also be used to administer actual midterm exams to students, with each student receiving a unique test that dynamically adapts to their proficiency level and reports their score immediately after the completion of the test.

In addition to giving students a realistic estimate of their current level of preparedness to take midterm exams, research has shown that engaging students in testing leads to better learning and long-term retention (i.e., the testing effect) compared to passive study methods commonly employed by students [7,8]. By using strategies such as rereading course notes or reviewing old homework problems to study, students can become overly familiar with the material and confuse familiarity with the surface features of problems with competence. This illusion of understanding is especially problematic for less prepared students because these students are less likely to test themselves using problems similar to those that will appear on midterm exams [7].

The benefits of the testing effect are thought to occur because retrieval attempts during testing facilitate deep processing of the material, strengthen pathways for correctly recalled information, and weaken pathways for incorrectly recalled information [9]. An additional benefit of testing appears to be test-potentiated learning, a term used to describe enhanced learning from studying that follows testing [10], even for new material encountered after initial testing [11,12]. When paired with correctness feedback, the testing effect has been found for items that were initially answered correctly [13], for items initially answered incorrectly during initial testing [14], and for untested but related items [15,16]. While much of the research in this area has focused on laboratory-based memory tasks [17,18], some studies have begun to explore the benefits of the testing effect in classroom settings [19,20], for complex tasks such as reading comprehension and inference tasks [21], and for problem solving in introductory STEM courses [22].

The present investigation reports the results for three studies that explore the reliability and validity of the CAT algorithm developed for two midterm exams in a large, calculus-based introductory mechanics course at the University of Illinois, as well as the usefulness of CAT to predict exam scores (experiments 1 and 2), and to serve as a diagnostic tool by providing students with information to guide their studying (experiment 3). In sum, our goals centered around investigating the benefits of CAT for students in an introductory physics course. More specifically, we asked (i) is testing using CAT reliable (i.e., performance is comparable across CAT administrations) and valid (i.e., the order of students' performances is maintained from CAT to P&P tests); (ii) does engaging in practice testing using CAT help students score higher on midterm exams when provided with diagnostic information; and (iii) can CAT tests accurately predict how students perform on the real exam (i.e., a student's score on the CAT test is similar to their score on the P&P test)?

In our studies, students volunteered to take CAT tests, and we compared the proficiency score provided by the CAT to their actual score on the two P&P midterm exams. We begin by discussing the use of classical test theory and item response theory (IRT) to construct the large item pool needed for CAT that covers two midterm exams. We then discuss some of the technical details of the test administration platform constructed to adaptively select test questions from a large item pool. Next, we describe three experiments that we performed, two examining the reliability and validity of the CAT algorithm, and one exploring

[1]The latent construct estimated by CAT is often referred to as "ability" in the CAT literature, however, we use the term proficiency to reflect our position that students' physics knowledge and problem-solving capabilities can evolve over time.

the diagnostic potential of CAT. We also explore the possibility for extending the benefits of CAT to an underexplored area by using the proficiency estimates from the CAT tests to predict midterm exam performance. We conclude with a discussion of what we learned from our experiments and provide some commentaries on the potential of using CAT for both diagnostic and summative purposes in large introductory science courses.

II. CONSTRUCTION OF ITEM POOLS, CAT ALGORITHM, AND DESIGN OF CAT PLATFORM

Design of CAT requires considerable knowledge of psychometric techniques (e.g. classical test theory and IRT) and of CAT-specific techniques (e.g., how to select items and monitor item usage). Our study had two experts in psychometrics and CAT design (H.A.K. and H.H.C.). In addition, we had programming expertise (GF) that allowed us to design a context-specific platform for administering the CAT for our experiments.

A. Procedure for constructing item pools

Two stages of analyses were conducted to build the item pools. The first stage consisted of analyses based on classical test theory [23,24] and used historical test data to examine the students' proficiency distributions and correlation between the item correctness and total score (i.e., point-biseral correlations). The second stage of analyses, based on three-parameter IRT [25,26], was conducted to evaluate whether the estimated item parameters have appropriate psychometric properties. These analyses indicated items that needed to be excluded from the item pool as well as suggested an appropriate item response model. The psychometric properties included the three item parameters (item difficulty, item discrimination, and a guessing parameter) as well as monotonicity, and dimensionality [25]. Each of these properties is discussed in turn. IRT models are essentially logistic regression models fit to each item that model the probability of a correct answer where the predictor variable is a student's value on the unobserved or latent construct (here denoted by the variable θ). In this study, we use the term physics proficiency to refer to the latent construct that is estimated from the response patterns. Item difficulty is the intercept, or location parameter, of the logistic regression model. The difficulty should fall in an appropriate range for the examinees, meaning that the item should not be too easy or too hard. If an item is too easy, in that most examinees got the item correct, then it is not very predictive of proficiency, and thus not useful. Similarly, if an item is too difficult, in that very few students got the item correct, then it provides no information about proficiency for less prepared students, and thus inappropriate for the item pool.

Item discrimination refers to the slope parameter of the logistic regression model. Highly discriminating items (i.e., large slopes) can distinguish between examinees with slightly larger or smaller values of θ. In other words, a highly discriminating item can "tease" apart examinees with different proficiencies with respect to the difficulty of the item. Items with low discrimination were excluded from item pools. We used a model with a "guessing" parameter because the exams in the course (and therefore the items used in CAT) assessed students with multiple-choice questions. Because students who may not know the answer are likely to guess, items with high guessing parameters (e.g., items with poor distractors) were excluded from the item pools.

Monotonicity of items means that the probability of getting an item correct increases with larger values of θ. In other words, students with greater proficiency have higher probabilities of answering an item correctly (i.e., the higher overall score a student receives, the higher the probability of answering an item correctly). This is a characteristic of a logistic regression model used as an IRT model. Items that did not display monotonicity were removed from the item pools.

Dimensionality in IRT is statistically defined as the minimum number of dimensions such that local independence holds [27,28]. In other words, the dimensionality of the model represents the minimum number of latent constructs such that for any given latent construct (i.e., θ), the students' probability of answering a question correctly is only based on their value of θ. Student proficiency can be viewed as either a unidimensional or a multidimensional latent construct. In a unidimensional model, answers to all of the questions are based on a single proficiency score. Figure 1(a) depicts a unidimensional model where $Q_1 - Q_p$ represent the questions given to an individual student, and θ represents physics proficiency. These questions can cover more than one topic area (e.g., kinematics, Newton's laws), however a single proficiency score is estimated. This means that this approach does not necessarily assume that the exam content is unidimensional, only that an individual's proficiency is consistent or homogeneous across topics within a single exam. In our

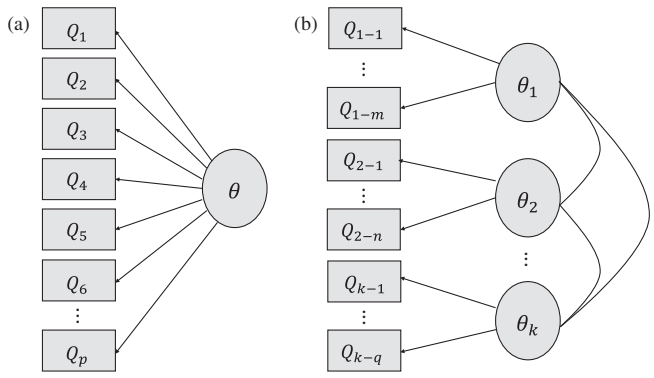

FIG. 1. (a) Unidimensional IRT model, (b) multidimensional IRT model.

context, this means that students who score highly on one content topic (e.g., conservation of momentum) are also likely to score highly on another topic found on the same exam (e.g., conservation of energy). The assumption of unidimensional physics proficiency is the common approach taken in summative P&P midterm exams given in introductory STEM courses since students are assigned a single grade representing their physics proficiency across several topics covered in the midterm, rather than a grade for each topic. In another example, the quantitative section of the Graduate Record Exam uses a unidimensional model, where the unidimensional θ represents overall math proficiency. The test is set up to select items in proportion to each subproficiency (e.g., algebra, data analysis) to ensure balanced coverage.

In contrast, a multidimensional model uses subsets of questions to estimate two or more proficiency scores. Figure 1(b) depicts a multidimensional model where $Q_{1-1} - Q_{1-m}$ represent the questions administered to an individual student from a certain topic (e.g., kinematics), while θ_1 represents the student's estimated proficiency on that topic. Similarly, $Q_{2-1} - Q_{2-n}$ represent the questions from a different topic (e.g., Newton's laws), while θ_2 represents the estimated proficiency on this second topic. The estimated proficiencies $(\theta_1 - \theta_k)$ can be combined to give an aggregate score. In addition, the proficiencies can be unequally weighted to give more importance to some topics than others. Importantly, if the dimensions (the estimated proficiencies $\theta_1 - \theta_k$) are highly correlated, the test essentially becomes unidimensional. For example, a multidimensional IRT model could provide subscores for kinematics and Newton's laws as well as a combined score that weighs Newton's laws more heavily than kinematics if one wished. Another example is the National Assessment of Educational Progress testing program, which estimates math proficiency as the (weighted) linear combination of five subproficiencies (number properties and operations, measurement, geometry, statistics, and algebra, respectively). To provide proficiency estimates for the subscales, tests that employ a multidimensional approach typically require a large number of questions (e.g., the complete National Assessment of Educational Progress exam consists of 230 questions).

In the current study, we adopted a unidimensional model for both practical and theoretical reasons, a decision supported by empirical evidence. Practically, one objective was to develop a testing platform to help less-prepared students study for upcoming midterm exams. Since the P&P midterm exams used in the course assumed a unidimensional model (only one overall score is ever computed), we wanted to develop a CAT that operated under similar assumptions. In addition, the CAT algorithm was constrained to select questions from the content areas such that they had a similar coverage with those on typical midterm exams given in the course. Theoretically, a unidimensional model is more efficient, assuming sufficient empirical evidence, both computationally and in terms of test length. We wanted to develop a tool that students might be likely to use, which meant that we needed to estimate students' proficiency using a relatively short test (a three-hour "practice test" would not be well received). In addition, we had a relatively limited pool of test questions from the subtopic pools. For example, note that the number of questions in a single administration of the National Assessment of Educational Progress exam is larger than our item pools.[2]

Empirically, unidimensionality was assessed by evaluating the fit of the model to the data using item-fit statistics under the unidimensional IRT model (i.e., three-parameter logistic model); note that we are not evaluating whether a unidimensional model is a better fit compared to multidimensional options; rather, we evaluated whether the hypothesized unidimensional model adequately describes the observed variance in the data and could be used for reporting physics proficiency scores. This is commonly done by checking absolute fit statistics [29,30]. In the present case, the statistical significance testing based on a χ^2 statistic produced p values of 0.375 on the second midterm exam item pool and 0.308 on the third midterm exam item pool, suggesting no model misfit. This indicates that the unidimensional three-parameter logistic model is able to serve as a latent variable model accounting for differences in the observed response patterns.

A general consensus in IRT is that for accurate estimation of the three-parameter logistic item response model, a sample size of 500 examinees is insufficient, whereas a sample size of 1000 is considered moderate [31–34]. A minimum sample size of 800 was required for items to be included in this study in order to construct item pools with a reasonably large number of items to achieve adequate precision. Two item pools were constructed based on 12 semesters of midterm exams when the enrollment was at least 800 students. One item pool was constructed for the second midterm exam; One hundred and eighty-seven questions were used out of 297 questions (topic coverage: work and energy and momentum). Another item pool was constructed for the third midterm exam; Two hundred and one questions were used out of 298 questions (topic coverage: rotational motion, angular momentum, and moment of inertia). Tests based on IRT that are used in CAT require a large database of prior multiple-choice questions that have been taken by a large number of examinees. The student performance statistics (e.g., number of students taking each item, the number of students selecting each choice for each item, etc.) for all multiple-choice exam

[2]It is possible that we either had a large enough pool of questions with sufficient psychometric power or one could have been obtained, to develop a multidimensional CAT. However, the tradeoff in efficiency, in both model complexity and test size, was not worth the increase in test length given our empirical results.

items administered over a dozen years have been maintained in a large data archive by the Physics Department, allowing for a detailed psychometric analysis of the old exam items to create the item pools for the CAT tests.

B. CAT platform and algorithm

The CAT tests were delivered through a web-based platform that presented a specified number of test items (17) on a computer screen according to the CAT algorithm used in this study and described below. The algorithm selected an item whose difficulty matched the examinee's estimated proficiency based on their performance on previous items, as well as the topics to be covered on the midterm exam (e.g., work, energy, and momentum). In general, if a student answered a question correctly, the next item was slightly more difficult and may come from any of the topics being sampled. Conversely, if a student answered a question incorrectly, the next item was slightly easier. However, the algorithm attempted to ensure that the topic coverage was representative of the midterm exam. After all 17 items were administered, a proficiency score was computed and linearly transformed into a percentage score (see Refs. [3,35,36] for more details of how items are selected for a CAT). In experiment 1, students were provided with a percentage score (i.e., 0%–100%) based on their estimated proficiency score. In experiment 2, students were not provided any feedback about their performance. In experiment 3, the CAT was promoted to students as a "diagnostic tool," and students were given the percentage score (i.e., 0%–100%) that was translated from their estimated proficiency score along with suggested topics to study based on their CAT performance (e.g., you got 1 out of 4 conservation of momentum problems correct).

The CAT algorithm initially assigns the first item assuming the student is of average proficiency (The average proficiency level was empirically determined by taking an average of the proficiency estimates calculated from archival data).[3] After the student answers the first item, the algorithm estimates the student's proficiency, then searches for the most informative item in the pool given the current estimate of the student's proficiency and the difficulty and discrimination of the remaining questions in the item pool by identifying the item that maximizes the Fisher information [2,26]. The CAT algorithm was designed to ensure that students would be presented with a representative sample of the content for each midterm exam. The procedure continues until the prespecified test length (number of questions) is met. During the CAT test, the selection of items is designed to minimize the standard error of the student's proficiency estimate as well as to accommodate the student's proficiency. Thus, the algorithm attempts to hone in quickly and efficiently on the student's "true" level of competence.

The adaptive nature of CAT requires dynamic evaluation of a student's level of competency in real time as they take an exam. This potentially allows for estimates of student proficiency to be made with fewer items than traditional P&P exams. However, there is a tradeoff between the number of items used in an assessment and the precision of the estimate of the test-taker's score on that assessment—the more items included on an assessment, the more precise the measure of the student's proficiency. In our case, the P&P midterm exams used between 24 and 26 questions.[4] For practical reasons, we decided to use a 17-question CAT test for this study to achieve balance between the desire for a shorter test and precise estimates of proficiency. We knew that students do not normally like to take one (or several) 25 question P&P practice test(s) to prepare for their midterms, so a 17-question CAT exam could be "sold" to student volunteers as a shorter test that would give them a reasonable prediction of their exam preparation and help them prepare for their midterms.

An additional factor that limited the predictive ability of the CAT algorithm was the possibility that students could earn partial credit on P&P exams. In the P&P midterm exams, approximately half of the questions have 5 choices and are worth 6 points. Students can select two choices in cases where they are not sure of the correct choice. If one of the two selected choices are correct, the student is given 3 points for the question (half credit). The other half of the questions have 3 choices and are worth 3 points. For these questions, students are only able to select one option and no partial credit is available. The CAT algorithm does not allow for partial credit because multiple answers to questions are not permitted by the algorithm. However, as mentioned above, the students' proficiency score is translated into a percentage score using historical midterm exam data where partial credit was available, thus the mean scores for groups of students should not differ.[5] Because of statistical variations for *individual students*, it is possible

[3]This means that the average proficiency level was empirically determined using historical data and is specific to the population at the university where this study was conducted.

[4]Midterm exams are constructed by the team of faculty teaching the course and they decide how many questions each midterm contains. The number of questions is a judgment call based on how many can be comfortably completed within the time allotted for the exam, which is 1.5 h for the course used in this study.

[5]The way this partial credit system is designed, the same mean should result with or without the partial credit. For example, suppose a student is not sure about the answers to two questions, but narrows down the possible answer to two choices for both questions. Suppose further that in case A the student chooses one of the answers in both cases, giving them a 50% probability of getting each correct. The expected payoff in case A is 6 points for the two questions. Suppose in case B that the student chooses both answers in both questions. The payoff in case B is also 6 points for the two questions. Thus statistically, the partial credit system is designed to yield the same mean.

that the midterm test varies slightly from the "true score" for a student due to partial credit, and hence the CAT could under- or overpredict *individual scores* resulting in lower correlations between the CAT tests and the P&P midterm exams.

III. EXPERIMENTS

We conducted three experiments in an introductory calculus-based mechanics course taken primarily by physics and engineering majors. The experiments were designed to evaluate the reliability and validity of the CAT algorithm, the predictive ability of CAT tests with respect to the actual P&P midterm exams administered in the course, and to investigate the potential to use CAT as an intervention aimed at helping low-performing students prepare for a midterm exam. The CAT algorithm estimates a latent physics proficiency score (as described above), and then translates this score to a percentage scale using a linear translation from the historical data described above. All analyses are conducted on this translated score so that all data (midterm exam and CAT scores) are on the same scale and are more interpretable when reported to students as part of the CAT feedback. Experiment 1 focused on the parallel-forms reliability of two CAT tests as well as exploring the potential for using CAT to predict midterm exam performance on an upcoming midterm exam. The criterion-related validity of CAT was also explored by examining the correlations between the CAT tests and the subsequent midterm exams. Because one of the major goals for developing the CAT platform was to help low-performing students prepare for midterm exams, it is important to establish that the CAT and the midterm exams are measuring similar constructs. In addition, because learning may have occurred between the CAT and the midterm exam in the first experiment, experiment 2 also investigated the criterion-related validity of a CAT test by examining its ability to predict midterm exam grades by delivering the CAT test the day *after* a midterm exam. Experiment 3 investigated the potential to use a CAT test as an intervention for low-performing students, to aid them in preparing for an upcoming exam.

A. Experiment 1

Experiment 1 explored the reliability of two CAT tests, as well as exploring the possibility for extending the benefits of CAT by using the proficiency estimates given by CAT to predict students' score on the third midterm exam. The reliability and predictive ability were examined by administering two CAT tests back to back the day before the real P&P third midterm exam. With this design, we evaluate how closely the two CAT test scores were to each other. In addition, the criterion-related validity of the CAT tests was assessed by examining the correlations between the CAT tests and the third midterm exam. Finally, the accuracy of the CAT tests for predicting the real midterm score was evaluated by conducting a sequential regression analysis. It should be noted that various factors can influence the ability of CAT (or any assessment) to predict an individual score on an exam. For example, differences in student motivation between the assessments (students are more motivated for course midterm exams), the length of the time interval between assessments (students are likely to continue their preparation until the day of the assessment), and the representativeness of the student sample. In this study, students had time after taking the CAT tests, and were likely motivated to use the information provided by the CAT tests to continue to prepare for their midterm exam. This means that it was expected that the scores on the CAT tests would be slightly lower than the scores on the midterm exam.

1. Participants

An email was sent out to all students ($N = 1229$) enrolled in the introductory calculus-based mechanics course a week before the third midterm exam. The email invited these students to participate in a study evaluating the usefulness of computer adaptive tests for diagnosing individual strengths and weaknesses to help students prepare for the upcoming exam. The email further explained that volunteers would take two CAT tests that would help them prepare for the third midterm exam. Students were told in the email that the CAT tests were designed to predict their grades on the third midterm exam, and that they could use the prediction to help them prepare for the third midterm exam. Thirty-four students volunteered and completed both CAT tests. The remaining 1195 students served as a comparison group.

2. Procedure

The 34 volunteers completed two CAT tests in a single three-hour session in the lab, one day before the third midterm exam. Because the CAT tests were taken back to back, and students received correctness feedback following the first CAT test, no questions from the first CAT test were repeated on the second CAT test for any individual student. The students completed the CAT tests under testing conditions similar to the conditions used on the real exam (e.g., formula sheets and scratch paper were provided, and students were able to use calculators, but no other resources). Students were given correctness feedback and copies of the questions from the CAT after completing each CAT test and were told their predicted score with the estimated standard errors. Students completed the third midterm P&P exam, consisting of 25 questions, the following evening.

3. Results

The average scores on the three midterm exams for the 1195 students in the comparison group who completed all

TABLE I. Means, standard errors of the mean, and intercorrelations for scores on exams and CATs. Note that $N = 34$ and that all correlations are significant at $p < 0.005$.

Measure	M	SE	1	2	3	4
1. Midterm 1	76.5	2.88	⋯			
2. Midterm 2	74.6	3.05	0.67	⋯		
3. Midterm 3	69.5	3.74	0.68	0.69	⋯	
4. CAT 1	58.2	2.57	0.48	0.60	0.78	⋯
5. CAT 2	58.2	2.52	0.53	0.57	0.68	0.57

three midterm exams were 79.0 ($sd = 14.5$), 78.3 ($sd = 14.9$), and 68.8 ($sd = 19.3$), respectively. The average scores on the three midterm exams for the 34 students who completed both CAT tests were 76.5 ($sd = 16.8$), 74.6 ($sd = 17.8$), and 69.5 ($sd = 21.8$), respectively. The correlations between the exams for the students in the comparison group were as follows: midterm 1 and 2 ($r = 0.63$), midterm 1 and 3 ($r = 0.65$), midterm 2 and 3 ($r = 0.70$). The means, standard errors, and Pearson correlations on the midterms and CAT tests for the 34 participants can be found in Table I. The standard errors for individual CAT scores ranged from 8% to 16%, with a mean of 10%.

(a) Do the two CAT tests give similar predictions?

To investigate the reliability of the CAT tests, three analyses were conducted. First, Pearson product-moment correlations were computed [37]. While no strict guidelines are universally accepted, Pearson correlations are often interpreted as follows; slight or poor correlations are indicated by values lower than 0.4, moderate correlations by values between 0.4 and 0.7, high correlations by values between 0.7 and 0.9, and very high correlations indicated by values higher than 0.9 [38]. The Pearson correlations suggest that the scores on the first CAT test were significantly, but moderately correlated with scores on the second CAT test ($r = 0.57$, $p < 0.001$). Second, since the predicted scores were based on IRT proficiency estimates a parallel-forms reliability analysis was conducted [39]. Values for the parallel-forms reliability coefficient range between -1.0 and 1.0, with 0 indicating no correlation, and either -1.0 or 1.0 indicating perfect correlation. The parallel-forms reliability coefficient suggests that the scores on the first CAT test were significantly, but moderately correlated with scores on the second CAT test ($\rho = 0.56$, $p < 0.001$). Third, intraclass correlations (ICC) [40–42] were calculated. Values for the ICC range between 0 and 1.0, with 0 indicating no reliability and 1.0 indicating perfect reliability. The ICC provide additional information about the reliability than Pearson correlations because they measure both the degree of correlation and degree of agreement between measurements. The ICC represent the proportion of variance in a set of scores that is attributable to the variance between individuals while the balance $(1 - \text{ICC})$ of variance is attributable to variance due to measurement error [43]. In other words, a test is reliable when more of the variance in the scores is the result of differences between individuals rather than differences between the two CAT test administrations for individual examinees, with poor reliability indicated by values lower than 0.5, moderate reliability by values between 0.5 and 0.75, good reliability by values between 0.75 and 0.9, and excellent reliability by values higher than 0.9 [44]. The ICC indicate moderate to good reliability for the two CAT tests ($\text{ICC}_{1,k} = 0.73$).

In addition to the correlation between the scores, the degree to which the CAT tests yielded the same scores was examined. A dependent-samples t test failed to indicate a significant difference between the two CAT test predictions, $t(33) = 0.003$, $p = 0.99$. Finally, a difference score was calculated by subtracting the first CAT test score from the second for the 34 students who completed both CAT tests (Fig. 2). A difference of zero indicates that both CAT tests gave the same prediction, a positive difference indicates a higher second prediction, while a negative difference indicates a lower second prediction. The mean of the difference scores was -0.01, indicating that, on average, the two CATs gave very close to the same prediction. However, the standard deviation was 13.8, and half of the students had CAT test scores that differed by more than 10 percentage points. In other words, the proficiency scores reported by the two CAT tests differed by at least one letter grade for half of the students.

(b) What is the evidence for the criterion-related validity of the CAT?

To answer this question, we conducted two analyses. First, we conducted Pearson product-moment correlations for the three course exams and the two CAT exams (Table I). The Pearson correlations suggest that the scores on the third midterm exam were strongly correlated with the scores on first CAT test ($r = 0.78$, $p < 0.001$), and moderately correlated with the scores on the second CAT test ($r = 0.68$, $p < 0.001$), and the second midterm exam ($r = 0.69$, $p < 0.001$).

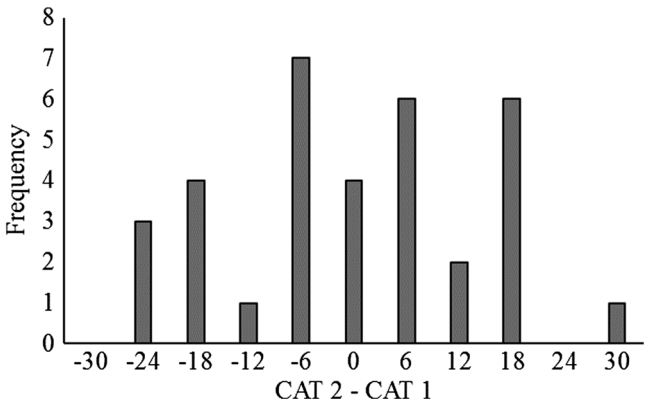

FIG. 2. Distribution of the difference scores between the two CAT administrations.

Second, two dependent-samples t tests were conducted to examine the accuracy of the predictions from the CAT test. As expected, the estimated scores reported by the CAT tests were lower than the score from the third midterm exam. The scores on each CAT test were about 11% lower on average than the third midterm exam, $t = 4.82$, $p < 0.001$; $t = 4.13$, $p < 0.001$, respectively.

(c) Can the two CAT tests predict the score on the third midterm exam?

To determine whether the CAT tests provided information about students' physics proficiency that could predict the score on the third midterm exam above the information available from the previous two midterm exams, a sequential multiple regression was conducted. In the first step the scores from the first two midterm exams were entered into the model. The first two midterms predicted the score on the third exam, $F(2, 31) = 19.99$, $p < 0.001$, $R^2 = 0.56$. In the second step the scores from the two CAT tests were entered, resulting in a significant increase in R^2 of 0.20, $F(2, 29) = 12.34$, $p < 0.001$. This indicates that the scores from the CAT tests explained an additional 20% of the variance in the third midterm exam scores above the previous two midterm exams. Running the scores from the two CAT tests first in the multilevel regression results in 69% of the variance explained, $F(2, 31) = 34.73$, $p < 0.001$, $R^2 = 0.69$.

(d) What factors might explain the observed underprediction?

The CAT underpredicted midterm exam 3 performance, meaning that the scores attained on the CAT were lower than the P&P scores (in this case, by 11%, on average). Although a slightly lower score was expected on the CAT tests due to additional studying, it is also possible that completing the CAT tests may have benefitted students on the third midterm exam. To see whether receiving a prediction by taking the CAT tests may have impacted performance, we calculated z scores for all 1229 students who completed all three midterm exams (this includes the 34 students who completed the CAT tests as well as the 1195 students in the comparison group). The comparison group had mean z score of zero for all three exams, while the CAT group (Fig. 3) had a negative mean z score for the first midterm exam (−0.13) and the second midterm exam (−0.23), but had a positive mean z score for the third midterm exam (0.04).

To examine how the groups compared on the third midterm exam while controlling for the z scores on the first two midterm exams, we conducted a one-way analysis of covariance (ANCOVA) with the third midterm z score as the response variable, group as the between-subjects variable, and the z scores from the first and second midterms as the covariates. Levene's test indicates that the assumption for homogeneity of variance was met, $F(1, 1227) = 1.36$, $p = 0.24$, however, a Shapiro-Wilk test indicates that the z scores on the third midterm exam were not normally distributed ($W = 0.97$, $p < 0.01$). However, the ANCOVA is robust to deviations from normality for unbalanced designs when a homogeneity of variance can be assumed and for sample sizes greater than 20 [45]. Since this study employed large samples and the distributions for the two groups were similarly negatively skewed, this test is appropriate for the data. To test the assumption of homogeneity of regression, we fit the model with the variate-covariate interactions. Because the interaction terms were not significant [exam 1 × group: $F(1, 1223) = 0.07$, $p = 0.79$; exam 2 × group: $F(1, 1223) = 0.30$, $p = 0.58$], the assumption of homogeneity of regression holds. The results (Table II) indicate that, unsurprisingly, students who score higher on both midterms exams 1 and 2 tend to score higher on the third midterm exam. In addition, although the effect of group is not significant and has a small effect size, it does approach significance ($p = 0.06$).

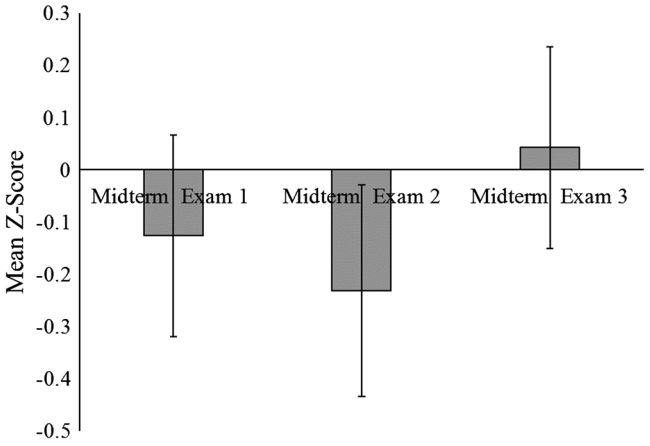

FIG. 3. Mean z score for students completing the CAT tests. Error bars represent the standard errors of the mean.

4. Discussion

The CAT tests demonstrated moderate reliability using Pearson correlations and parallel-forms reliability analysis, while ICCs suggest slightly stronger reliability. More

TABLE II. One-way ANCOVA for the effects of semester and student ability (proficiency) on midterm exam 3 scores. Note that type III sum of squares and mean squares, and semipartial eta square are reported.

Variable and source	df	SS	MS	F (1, 1225)	p	η^2
Group	1	1.57	1.57	3.50	0.06	<0.01
Midterm 1 z score	1	88.43	88.43	197.70	<0.001	0.07
Midterm 2 z score	1	171.68	171.68	383.83	<0.001	0.14
Within	1225	547.91	0.45			

specifically, half of the students received predictions that differed by more than 10 percentage points from their actual midterm score. One possible reason for this is that the CAT tests consisted of only 17 questions. The inclusion of more questions may lead to more stable proficiency estimates. Another potential reason for lower than desired reliability is that internal reliability measures assume that physics proficiency is homogeneous between the content areas of a given CAT test (e.g., conservation of momentum, work and energy). This means that the assumption that students' latent physics proficiency is homogeneous across content areas may be too strong. Future investigations of CAT could explore potential increases in reliability by using a multidimensional approach.

The scores on the CAT tests also were strongly correlated with the scores on the third midterm exam suggesting that the CAT tests exhibited good criterion-related validity. However, both CAT test scores were on average 11% lower than the scores on the third midterm exam. Several factors could have contributed to this underprediction, including lower than optimal reliability of the CAT tests and midterm exams. Alternatively, this difference in scores may indicate that the CAT tests and the midterm exams estimate student proficiency differently.

Another possibility for the difference in scores is that students who took the CAT tests may have been motivated by lower than desired CAT scores and engaged in additional studying. However, this conjecture is unclear given the results of the ANCOVA (also see the results from experiment 3). Students who took the CAT tests did not demonstrate statistically significantly higher z scores on the third midterm exam after controlling for z scores on the first two midterm exams ($p = 0.06$). Prior research indicates that engaging in testing facilitates learning [9,19]. However, we may have failed to find a significant difference here because of the limited amount of time between the CAT tests and the midterm exam. This may indicate that taking a practice exam one day before a midterm exam may not be sufficient time for students to fully realize the benefits of test-potentiated learning. Experiment 3 investigates the potential testing effects from engaging in CAT by having students complete the CAT tests earlier and on two separate occasions to maximize the use of feedback and test-potentiated learning.

The evidence concerning the ability of CAT tests to predict student scores on the exam is mixed. The CAT tests explained a significant amount of variance in the third midterm exam over that explained by the first two midterm exams alone, suggesting that CAT provides information about students' current proficiency beyond that available from the other midterm exams. However, the large standard deviation (13.8) of the difference score between the third midterm exam and the CAT tests indicates that predicting *individual* student scores is less than optimal. The large variance in predictions could occur because 17 questions may not be a large enough sample to yield completely accurate predictions. Another factor that may contribute to the large variance in predictions is the fact that the students could earn partial credit on the midterm exams, although the mean should not be affected by partial credit.

B. Experiment 2

In experiment 1, it was possible, maybe even probable, that students who volunteered for the CAT tests did a considerable amount of studying the evening before the third midterm exam, especially given the low scores reported to them that afternoon on the CAT tests. This possible cramming could have been responsible for the observed underprediction in the two CAT back-to-back test administrations in experiment 1. In experiment 2, a single CAT test was administered the day *after* the second midterm exam, which minimizes the possibility of any additional studying occurring. This design provides a more faithful measure of the CAT test's criterion-related validity as measured by the ability of the CAT test to predict performance on the second midterm exam.

1. Participants

An email was sent out to all students enrolled in the introductory calculus-based mechanics course inviting them to participate in the study. Students were told that the study would evaluate the predictive ability of a CAT test against performance on the real second midterm exam (the second midterm contained 24 P&P multiple choice questions). Because there was no motivation for students to take another test on the day after the second midterm exam, students were offered $10 as an incentive to participate. A total of 76 students volunteered for the experiment. The students had a mean first midterm exam score of 78.7% and represented a range of first midterm exam scores ($sd = 17.0\%$, range: 30%, 100%).

2. Procedure

The 76 volunteers came to the computer-resource lab in the physics department on the day immediately following the second midterm exam and took a CAT test covering the same material as the second midterm exam. Students were not given any feedback about their performance for this experiment. Because students had just taken the real exam the night before, it was hypothesized that there was little forgetting of the material, and that no additional studying took place between the second midterm exam and the CAT test.

3. Results

The class average score on the first midterm exam was 76.7 ($sd = 18.4$), and the class average score on the second midterm exam was 72.9 ($sd = 16.6$). The means, standard errors, and Pearson correlations for the two midterm exams

TABLE III. Means, standard errors of the mean, and intercorrelations for scores on exams and CAT. Note that $N = 76$, all correlations are significant at $p < 0.001$.

Measure	*M*	*SE*	1	2
1. Midterm exam 1	78.7	1.95	⋯	
2. Midterm exam 2	76.7	1.80	0.74	⋯
3. CAT	76.8	1.60	0.60	0.58

and the CAT test for the 76 participants can be found in Table III. The standard errors for individual CAT scores ranged from 4% to 7%, with a mean of 5%.

(a) Is the proficiency estimate from the CAT test similar to the score on the midterm exam?

Since one of the major goals for developing the CAT platform was to help low-performing students prepare for midterm exams by providing feedback, it is important to establish that the CAT and the midterm exams measure similar constructs (i.e., physics proficiency). To examine the criterion-related validity of the CAT, two analyses were conducted. First, the Pearson correlations (Table III) suggest that the scores on the CAT test were significantly, but moderately correlated with scores on the second midterm exam ($r = 0.58$, $p < 0.001$). However, when the reliabilities of the assessments are less than perfect the correlation between the assessments is necessarily reduced (validity attenuation) [46]. To examine the estimated correlation between the CAT and the midterm exam that would occur if the ability (proficiency) scores were measured with perfect reliability we calculated the internal reliability of the midterm exam using Cronbach's α and found that the internal reliability of the second midterm exam was moderate ($\alpha = 0.77$). From experiment 1 we found that the parallel-forms reliability of the CAT test was also moderate (parallel forms $r = 0.56$). We used Eq. (1) in Ref. [46] to correct for the moderate reliability of the two exams and found a much stronger correlation ($r = 0.87$, $p < 0.001$). This provides evidence that the CAT test could potentially measure physics proficiency in similar to traditional P&P midterm exams given in the course, if the reliability of the CAT were improved.

Second, a difference score was calculated by subtracting the second midterm exam score from the score on the CAT test. A positive difference indicates that the CAT overpredicted the exam score, while a negative difference indicates that the CAT underpredicted the exam score. A Shapiro-Wilk test indicates that the difference scores were relatively normally distributed ($W = 0.99$, $p = 0.69$) with a mean of 0.15, and a standard deviation of 13.7. Not surprisingly, a dependent-samples t test failed to detect a difference between the CAT and the second midterm exam, on average, $t(75) = 0.10$, $p = 0.92$. However, only 59% of the study participants' CAT test scores were within 10 percentage points of their score on the second midterm. In other words, while the averages were similar (see Table II), the CAT test score and the midterm exam score for any individual student may differ by a letter grade or more.

(b) Can the CAT test predict the score on the second midterm exam?

To determine whether the information that the CAT test provided about students' physics proficiency could predict the score on the second midterm exam, above the information available from the previous course exam, a sequential multiple regression was conducted. In the first step, the first midterm exam was entered into the model. The first midterm exam predicted the score on the second midterm exam, $F(1, 74) = 87.95$, $p < 0.001$, $R^2 = 0.54$. In the second step the CAT test score was entered, resulting in a small but statistically significant increase in R^2 of 0.03, $F(1, 73) = 4.93$, $p = 0.03$. This indicates that the score from the CAT test explained an additional 3% of the variance in the second midterm exam above and beyond the information already available from the first midterm exam. Running the CAT test first in the regression results in 33% of the variance explained, $F(1, 74) = 36.75$, $p < 0.001$, $R^2 = 0.33$.

4. Discussion

The scores on the CAT test demonstrated moderate correlations with the scores on the P&P second midterm exam. In contrast to experiment 1, students completing the CAT test also received the same score, on average, as the midterm exam. This finding provides some evidence to establish the criterion-related validity of the CAT tests. However, the large standard deviation (13.7) in the difference score between the second midterm exam and the CAT test indicates that predicting *individual* student scores is less than optimal −41% of the study participants were either over- or underpredicted by more than 10%. A similar finding was found in experiment 1 when comparing the scores from the two CAT tests. The variation in the accuracy of the individual predictions suggest that the CAT tests and the midterm exams may estimate student proficiency in slightly different ways. Since the proficiency estimates are dynamically estimated after each additional question, the proficiency estimate given by the CAT test would very likely have improved if more than 17 questions had been used in the CAT test. Other factors could have contributed to the CAT test's difficulty in predicting *individual* performance on the second midterm exam. For example, the ability for students to earn partial credit on the midterm exam add additional error variance into the scores on the midterm exams. In other words, the midterm test score may differ from the "true score" for any individual student. This added variance, combined with the uncertainty in the CAT test scores may have contributed to the relatively large standard deviation in difference scores.

In addition to exploring the validity of the proficiency estimates provided by CAT, the results provide additional mixed evidence for the ability of CAT to provide

information allowing for predictions of student scores on the midterm exam. The CAT test explained a small portion of the variance in the second midterm exam above the score on the first midterm exam. However, this additional variance was much smaller in experiment 2 (3%) as compared to experiment 1 (20%). One reason for this difference is that more information about students' proficiency was available in experiment 1 (2 CAT tests) than in experiment 2 (1 CAT test). Given the moderate reliability of the CAT, taking two CAT tests may provide a more reliable estimate of students' "true proficiency." In addition, the full multiple regression in experiment 2–1 CAT test, 2 Midterm exams—explained less variance (57%) than the full multiple regression in experiment 1–2 CAT tests, 1 midterm exam—(76%). This result seems to indicate that the results from a single, 17-item, CAT test may be less useful for instructors than the results from either multiple or longer CATs tests.

C. Experiment 3

One motivation for developing the CAT platform was to provide feedback to low-performing students about their preparedness for the midterm examination to combat the illusion of understanding suggested by the Dunning-Kruger effect reviewed earlier. In addition, by engaging students in repeated testing before a midterm exam we hoped to take advantage of test-potentiated learning and the testing effect. The results from experiment 1 suggest that students may improve their midterm exam scores by using the CAT test as a study tool. In this experiment, low-performing students were invited to complete two CAT tests—the first administered three to five days before the second midterm exam, and the second administered one or two days before the second midterm exam. This design allowed us to explore the usefulness of CAT as an intervention aimed at helping low-performing students prepare for a midterm exam.

1. Participants

An email was sent to 292 students enrolled in the introductory mechanics course who had performed in the bottom third on the first midterm exam (that is, scored <76%). The email invited these students to participate in a study evaluating the usefulness of a CAT test for diagnosing weaknesses. The email further explained that volunteers would take two CAT tests that would help them prepare for the second midterm exam. Students were told that the CAT tests were designed to predict their grades on the midterm exam, and that they would be given specific feedback on which topics they did poorly on, so they could study the topics in detail before the actual exam. Thirty-three students volunteered and took the first CAT test. Of those, 25 returned to take the second CAT test. The 259 students who did not volunteer to take the CAT served as the comparison group for this study.

2. Procedure

The 33 volunteers took the first CAT test on their computers at home or in their dorms the weekend before their second midterm exam was scheduled. Afterwards they received an email telling them the score predicted by the CAT, as well as an error range (typically between ±4% and ±8%), together with topics that they did poorly on and recommended for further study (e.g., you got three out of four conservation of mechanical energy questions correct). Students came to the computer-resource room in the physics department one to two days before the second midterm exam was scheduled to take a second CAT test. Twenty-five students returned to take the second CAT test. The students received the same feedback as when they took the initial CAT test. Since this experiment was interested in the effect of CAT as an intervention, only the data from the 25 students who completed both CAT tests (CAT group) were analyzed. The predicted scores from the CAT tests were compared against each other, and against the scores on the second midterm exam.

3. Results

(a) Do students who complete two CAT tests score higher on the second midterm exam?

Means and standard deviations for the CAT group and comparison group can be found in Table IV. To assess whether taking the CAT tests help students to prepare for the exam as suggested by the testing effect, a 2×2 (Exam × Group) mixed ANOVA with exam score as the repeated measure and group as the between-subjects variable was conducted to analyze how the two groups differed in performance on the exams. Exam was significant, $F(1, 282) = 6.87$, $p < 0.01$, indicating that the average z score on exam 2 was higher than the z score for exam 1. However, while the main effect of *group* was not significant, $F(1, 282) = 0.37$, $p = 0.54$, there was a significant interaction between exam z score and group, $F(1, 282) = 3.79$, $p = 0.05$. The result of this analysis indicates that a different pattern of exam performance was found for the two groups, with the students who completed the CAT tests demonstrating greater improvement than those in the comparison group. Follow-up dependent t tests indicated that students who completed the CAT tests

TABLE IV. Means and standard errors of the mean for scores on course exams and CATs. Note that the CAT group is $N = 25$, the comparison group $N = 259$.

	CAT group		Comparison group	
Exam	*M*	*SE*	*M*	*SE*
Midterm exam 1	59.5	2.44	60.7	0.84
Midterm exam 2	66.2	3.00	61.7	0.93
CAT 1	67.3	2.38	⋯	⋯
CAT 2	69.1	2.42	⋯	⋯

scored higher on the second midterm exam with a Cohen's d effect size for repeated measures (d_z) indicating a moderate effect size representing an improvement of more than six percentage points, or two-thirds of a grade point, $t(24) = 2.79$, $p = 0.01$, $d_z = 0.56$, whereas those in the comparison group did not score statistically significantly higher on the second midterm exam, $t(258) = 1.13$, $p = 0.26$, $d_z = 0.07$.

To evaluate the hypothesis from experiment 1 that the underprediction was due, in part, to students engaging in extra studying between the CAT tests and the midterm exam, two dependent-samples t tests were conducted to examine the accuracy of the predictions from the CAT test. Shapiro-Wilk tests indicate that the difference scores were normally distributed for both the first ($W = 0.99$, $p = 0.31$) and second ($W = 0.97$, $p = 0.61$) CAT tests. The dependent-samples t test failed to detect a difference between the first CAT test and the second midterm exam, on average, $t(24) = 0.40$, $p = 0.70$ and between the second CAT test and the second midterm exam, $t(24) = 0.90$, $p = 0.37$. However, a large number of individual estimates from the CAT tests differed from the scores on the second midterm by more than 10 percentage points (40% and 48% respectively).

4. Discussion

It appears that students who took both CAT assessments made significantly larger improvement from first to second midterm exam compared to the comparison group (i.e., the lower-third pool of students who did not volunteer to participate in the study). Although we cannot attribute causality to this finding because the study lacked a true control group, we offer some possible reasons why the students taking the CAT tests may have improved their performance relative to the comparison group. It is possible that the information about the students' preparedness for the exam motivated them to engage in additional studying. Students were also provided information about the content areas that they were weakest in, which may have helped students to focus their studying.

Left to their own devices, students typically select study strategies that tend to be passive and focus on encoding processes such as rereading, reviewing notes, or homework [47]. It may be that the use of testing as a study strategy may have allowed students to take advantage of test-potentiated learning. Testing, as a study strategy, helps students learn by improving the encoding of new information, and by helping students calibrate their other study strategies [48]. The feedback about exam readiness and individual strengths or weaknesses provided to students from the CAT tests likely motivated them to engage in additional, and more targeted studying for the second midterm exam.

An alternate explanation is that the distributed nature of the CAT administrations encouraged students to engage in more distributed practice, a strategy shown to improve long-term retention more than massed practice [15,49]. Two-thirds of undergraduate college students report using cramming as a primary study strategy and over one-half report the tendency to study in one session immediately before a test, that is, they use a massed form of practice [7]. Another possibility is that the effect found in this study is due to the higher motivation of the students who volunteered for the study. While the higher motivation was not evident on the scores for the first midterm exam, it could be that the lower-than-expected scores on the first midterm exam was one of the motivating factors in the improvement on the second midterm exam.

The ability of the CAT tests to estimate students' proficiency in the same way as the midterm exams is unclear. In contrast to the findings from experiment 1, the average proficiency estimate from the CAT tests were the same as the midterm exam. One interpretation of this finding is that the significant interaction found by the ANOVA is the result of motivation differences rather than from engaging students in testing. However, this interpretation implies that no learning occurred in the week before the exam. This interpretation does not seem likely, given that students tend to employ cramming as a test preparation strategy [7]. An alternative interpretation is that CAT tests and midterm exams estimate student proficiency slightly differently. This interpretation is also consistent with the differences in *individual* proficiency estimates between the CAT tests and the Midterm exam.

IV. GENERAL DISCUSSION

This study explored the potential of computer adaptive testing for diagnosing weaknesses to provide specific feedback for students to use in studying and for predicting midterm scores in exams. This study is one of the first attempts to use CAT in a large introductory STEM course, in this case a calculus-based mechanics course enrolling over 1000 students. Experiment 1 explored the reliability and validity of the proficiency estimates reported by the CAT tests. The correlations demonstrate that the CAT algorithm displays moderate reliability and validity. Although the CAT algorithm may be appropriate for exam preparation, the CAT algorithm needs further investigation before being implemented as a summative exam in introductory courses.

In addition to exploring the reliability and validity of the CAT algorithm, students completing two CAT tests the day before a course exam demonstrated higher, though nonsignificant ($p = 0.06$), z scores on the third midterm exam after completing two formative CAT tests. The nonsignificant result indicates that the CAT exams did not help students perform better when given the day before the third midterm exam. This might suggest that students need more than one day to take advantage of test-potentiated learning benefits from testing. In addition, the CAT tests underpredicted performance on the third midterm exam on average. Finally, while the CAT algorithm was not developed

to predict future exam scores, there is some evidence for the predictive ability of CAT. The average CAT test score explained an additional 20% of the variance in the scores on the third midterm exam. However, the large variance in the difference scores indicates room for improvement in predicting individual exam scores. Future research should investigate improving the quality of proficiency estimates by exploring the number of questions needed for more accurate predictions, incorporating more efficient item selection methods, improving the methods for balancing content topics, and by utilizing the student's response time to help calculate a proficiency score [50].

In the second experiment, in which the CAT was administered the day after the second midterm exam, the average on the CAT exam was almost identical to the second midterm average. In addition to the correlations between the scores on the midterm and the CAT test, this provides additional evidence for the validity of the scores given by CAT. However, the CAT test's ability to predict scores for individual students was less than optimal given the large standard deviation of CAT exam scores. It should be noted that the CAT tests used in all our studies consisted of 17 questions, while midterm exams generally have 24–25 questions, so one way of improving the prediction of CAT tests would be to increase the number of items. There is a tradeoff between exam efficiency and the precision of the proficiency estimate.

The third experiment explored whether CAT tests administered a few days before a midterm exam could serve as a diagnostic tool for underperforming students. In that experiment, underperforming students from the first midterm exam who volunteered to take two CAT tests were given the predicted scores as well as information on the topics that they did poorly on to help them target their studying for the midterm exam. Although the study group and the comparison group were not statistically different in their average performance on the first midterm exam, the group who took two CAT tests scored about seven percentage points higher on the second midterm exam compared to the first midterm exam, while the comparison group only scored about 1 percentage point higher on the second midterm exam compared to the first midterm exam. This finding suggests that CAT could be a promising strategy to help underperforming students prepare for exams. This is perhaps not surprising in view of prior research indicating that underperforming students walk into exams overconfident and underprepared since their study habits (which usually do not include self-testing) is geared to provide familiarity with the material and not necessarily competence with the material [7,47,51].

The results from the three experiments appear somewhat paradoxical. The CAT tests underpredicted the scores on the midterm exams in experiment 1, but this was not the case in experiment 3. The original hypothesis that the underprediction was the result of students engaging in additional studying before the midterm exam appears unlikely because we would expect the first CAT test in Experiment 3 to also underpredict the midterm exam score. An alternative conjecture is that the underprediction observed in experiment 1 is related to the ways in which the CAT tests and the midterm exams estimate student proficiency. Traditionally P&P exams estimate student proficiency by computing the percentage of available points an individual student earns on an exam. This method typically does not use empirical data (e.g., difficulty, discrimination, etc.) when constructing the exams or computing the points students earn. In our context, instructors generally make up P&P exams from experience with the goal of writing a test with a mean of approximately 75% (sometimes successfully, sometimes not). Conversely, the CAT algorithm estimates proficiency using a latent (unobserved) score that is computed using empirical data from questions completed by previous students as described above. A difference in proficiency estimation may also explain the difference of 10 percentage points between proficiency estimates given by the CAT tests and the midterm exams experienced by approximately 40% of the students in all three experiments. Further research is needed to explore the consistency of proficiency estimates obtained across both testing platforms.

Is CAT a viable option for large introductory STEM courses either as a diagnostic tool to help underperforming students prepare for exams or as an actual test-administration tool instead of using paper-and-pencil exams? We believe that it is unlikely that CAT is a viable option for adoption in most STEM departments due to the extensive resources needed to develop and implement CAT and the relatively modest benefits in student outcomes seen in this study. In terms of resources, developing the CAT system requires considerable start-up costs. To construct a large item pool, a large number of examinees are initially needed (greater than 800) to evaluate whether the estimated parameters for a test question have appropriate psychometric properties (e.g., discrimination, difficulty level) before including them into the final item pool. However, for new exam questions, the item parameters can be calibrated during subsequent CAT administrations, which can save time and money by reducing the number of examinees (approximately 100–200) needed before including items into the final item pool. Furthermore, substantial psychometric expertise and coding expertise are needed to prepare an adequate item pool and to develop the web-based CAT platform, both of which may not be available in STEM departments. However, once built, the database and web-delivery platform can be used repeatedly at minimal additional cost. In addition, if CAT is primarily used to diagnose students' preparation for an upcoming exam in order to help them study, then there is less concern about security breaches. Finally, more advanced techniques, such as online calibration [52–54], could be embedded into the platform, thereby reducing the cost for expanding or replenishing the item pools.

Despite these obstacles, there may be benefits for those STEM departments with the resources to develop and implement CAT both in terms of helping students learn and as a means of assessing students. We had hoped that one benefit of CAT would be to provide efficient estimates of the examinee's proficiency level with fewer items than traditional 24–27 question practice exams. Moderate parallel-forms reliability was noted for the CAT algorithm implemented in this study. This suggests that the expected efficiency benefit was not realized in this implementation. As such, more research is needed to improve the reliability before implementing CAT as a summative course assessment tool. However, CAT might be useful to students as a formative assessment tool. By providing students with feedback about their preparedness for a midterm exam, along with information about the content areas and skills in which they are weakest, CAT might serve to motivate students to do additional studying, helping them to better prepare for exams. As a diagnostic tool, CAT tests serve the important function of providing underperforming students with a realistic estimate of their preparation for exams prior to taking them, as well as providing students with specific information to help them target their studying. This would help mitigate what has been referred to as the "double curse" of underprepared students:

> *"In many significant social and intellectual domains, the skills necessary to recognize competence are extremely close if not identical to those needed to produce competent responses. … Thus, incompetent individuals suffer a double curse: Their deficits cause them to make errors and also prevent them from gaining insight into their errors. Several studies have now shown that incompetent individuals (i.e., those performing poorly relative to their peers) fail to show much insight into just how deficient their performance is (Kruger & Dunning, 1999). … College students scoring in the bottom 25% on a course exam walked out of the exam room thinking that they outperformed a majority of their peers (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Compared with good students, poor students less successfully identify which specific questions they have gotten right on an exam and which they have gotten wrong (Sinkavich, 1995)." (Ref. [51], pp. 73–74.)*

Having a convenient CAT platform that delivers practice exams to students and provides immediate score predictions along with diagnostic information is a way of injecting some realism into underperforming students' thinking. Our experience in teaching large introductory courses, from talking to many students over the years who complain that they studied very hard and define themselves as "A students" yet did poorly in a physics exam, suggests that they do not self-evaluate their preparation for an exam by taking previous years' exams made available to them in the course web site. One motivation for developing the CAT platform was to study the potential benefits of providing less prepared students with problem-solving practice and realistic feedback with less time commitment. The hypothesis being that students might be more likely to use testing as a study strategy if it could be accomplished in less time. However, a limitation of this study was that we did not assess whether students would use the CAT on their own, or whether students will use CAT more frequently than traditional P&P practice exams. These are open empirical questions that need to be addressed by future research. An additional limitation was the relatively low volunteer rate from the courses employed in this study. Students may have been unlikely to volunteer for a study the week of a midterm exam, since undergraduate students, especially less prepared students, tend to view testing as a method to measure current learning rather than as a mechanism to enhance learning [55,56]. The participation rate prevented the use of control groups resulting in an unbalanced quasiexperimental design. While statistical controls were implemented (e.g., ANCOVA in experiment 1) future research should look to replicate and advance our findings.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grant No. DRL 1252389. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Prof. Carolyn Anderson for her suggestions on explaining the psychometric properties of computer adaptive testing.

[1] H. Wainer and R. J. Mislevy, Item response theory, item calibration, and proficiency estimation, in *Computerized Adaptive Testing: A Primer*, edited by H. Wainer (Lawrence Erlbaum Associates, New Jersey, 1990), pp. 65–102.

[2] D. J. Weiss, Improving measurement quality and efficiency with adaptive testing, Appl. Psychol. Meas. **6**, 473 (1982).

[3] H.-H. Chang, Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond, in *Handbook of Quantitative Methods for the Social Sciences*, edited by D. Kaplan (Sage, Thousand Oaks, CA, 2004), p. 117–133.

[4] D. Dunning, C. Heath, and J. M. Suls, Suls, Flawed self-assessment: Implications for health, education,

and the workplace, Psychol. Sci. Publ. Interest **5**, 69 (2004).
[5] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, J. Personality Social Psych. **77**, 1121 (1999).
[6] N. Falchikov and D. Boud, Student self-assessment in higher education: A Meta-Analysis, Rev. Educ. Res. **59**, 395 (1989).
[7] M. K. Hartwig and J. Dunlosky, Study strategies of college students: Are self-testing and scheduling related to achievement?, Psychon. Bull. Rev. **19**, 126 (2012).
[8] H. L. Roediger III. and A. C. Butler, The critical role of retrieval practice in long-term retention, Trends Cognit. Sci. **15**, 20 (2011).
[9] N. Kornell, M. J. Hays, and R. A. Bjork, Unsuccessful retrieval attempts enhance subsequent learning, J. Exp. Psychol. Learn. Mem. Cogn. **35**, 989 (2009).
[10] N. C. Soderstrom and R. A. Bjork, Testing facilitates the regulation of subsequent study time, J. Memory Language **73**, 99 (2014).
[11] E. L. Bjork and B. C. Storm, Retrieval experience as a modifier of future encoding: Another test effect, J. Exp. Psychol. Learn. Mem. Cogn. **37**, 1113 (2011).
[12] K. T. Wissman, K. A. Rawson, and M. A. Pyc, The interim test effect: Testing prior material can facilitate the learning of new material, Psychon. Bull. Rev. **18**, 1140 (2011).
[13] A. C. Butler, J. D. Karpicke, and H. L. Roediger III., Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses, J. Exp. Psychol. Learn. Mem. Cogn. **34**, 918 (2008).
[14] L. E. Richland, L. S. Kao, and N. Kornell, Can unsuccessful tests enhance learning?, in *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX, 2008), p. 2338.
[15] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. K. Kang, and H. Pashler, Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction, Educ. Psychol. Rev. **24**, 369 (2012).
[16] J. L. Little, E. L. Bjork, R. A. Bjork, and G. Angello, Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting, Psychol. Sci. **23**, 1337 (2012).
[17] S. H. K. Kang, T. H. Gollan, and H. Pashler, Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning, Psychon. Bull. Rev. **20**, 1259 (2013).
[18] H. L. Roediger III. and J. Karpicke, The power of testing memory: Basic research and implications for educational practice, Perspectives Psych. Sci. **1**, 181 (2006).
[19] M. A. McDaniel, K. M. Wildman, and J. L. Anderson, Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study, J. Appl. Res. Memory Cogn. **1**, 18 (2012).
[20] K. B. McDermott, P. K. Agarwal, L. D'Antonio, H. L. Roediger III., and M. A. McDaniel, Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes, J. Exp. Psychol. Appl. **20**, 3 (2014).
[21] M. A. McDaniel, D. C. Howard, and G. O. Einstein, The read-recite-review study strategy: Effective and portable, Psychol. Sci. **20**, 516 (2009).
[22] J. W. Morphew, G. Hermann, M. Silva-Sohn, and M. West, The testing effect and problem solving: Effect of more frequent assessment on student learning in a university STEM course (to be published).
[23] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory* (Waveland Press, Inc., Long Grove, IL, 1979).
[24] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Cenage Learning, Mason, OH, 2006).
[25] A. Birnbaum, Theories of mental test scores, in *Statistical Theories of Mental Test Scores*, edited by F. M. Lord and M. R. Novick (Addison-Wesley, Reading, 1968), pp. 397–479.
[26] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Lawrence Erlbaum Associates, Mahwah, NJ, 1980).
[27] W. F. Stout, A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation, Psychometrika **55**, 293 (1990).
[28] J. Zhang and W. Stout, The theoretical detect index of dimensionality and its application to approximate simple structure, Psychometrika **64**, 213 (1999).
[29] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika **37**, 29 (1972).
[30] W. M. Yen, Using simulation results to choose a latent trait model, Appl. Psychol. Meas. **5**, 245 (1981).
[31] F. B. Baker, An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure, Appl. Psychol. Meas. **22**, 153 (1998).
[32] C. L. Hulin, R. I. Lissak, and F. Drasgow, Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study, Appl. Psychol. Meas. **6**, 249 (1982).
[33] S. P. Reise and J. Yu, Parameter recovery in the graded response model using MULTILOG, J. Educ. Measure. **27**, 133 (1990).
[34] K. L. Tang, W. D. Way, and P. A. TOEFL Educational Testing Service, Princeton, NJ, Technical Report TR-7, 1993.
[35] H-H. Chang, Psychometrics behind computerized adaptive testing, Psychometrika **80**, 1 (2015).
[36] H.-H. Chang and Z. Ying, A global information approach to computerized adaptive testing, Appl. Psychol. Meas. **20**, 213 (1996).
[37] W. F. Holmes, B. F. French, and J. C. Immekeus, *Applied Psychometrics Using SAS* (Information Age Publishing, Charlotte, NC, 2014).
[38] R. C. Sprinthall, *Basic Statistical Analysis,* 8th ed. (Pearson Education, Boston, MA, 2007).
[39] S. Kim, A note on the reliability coefficients for item response model-based ability estimates, Psychometrika **77**, 153 (2012).
[40] M. Bedard, N. J. Martin, P. Krueger, and K. Brazil, Assessing reproducibility of data obtained with instruments based on continuous measurements, Experimental Aging Research **26**, 353 (2000).

[41] W. Kroll, A note on the coefficient of intraclass correlation as an estimate of reliability, Research quarterly for exercise and sport **33**, 313 (1962).
[42] P. E. Shrout and J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability, Psychiatric bulletin **86**, 420 (1979).
[43] J. P. Weir, Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM, J. Strength Conditioning Res. **19**, 231 (2005).
[44] T. K. Koo and M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropractic Medicine **15**, 155 (2016).
[45] G. Keppel and T. D. Wickens, *Design and Analysis: A Researcher's Handbook*, 4th ed. (Pearson Education, Upper Saddle River, NJ, 2004).
[46] P. M. Munchinsky, The correction for attenuation, Educ. Psychol. Meas. **56**, 63 (1996).
[47] J. D. Karpicke, A. C. Butler, and H. L. Roediger III., Metacognitive strategies in student learning: Do students practice retrieval when they study on their own?, Memory **17**, 471 (2009).
[48] J. Fernandez and E. Jamet, Extending the testing effect to self-regulated learning, Metacognition Learn. **12**, 131 (2017).
[49] K. A. Rawson, J. Dunlosky, and S. M. Sciartelli, The power of successive relearning: Improving performance on course exams and long-term retention, Educ. Psychol. Rev. **25**, 523 (2013).
[50] E. M. Choe, J. L. Kern, and H.-H. Chang, Optimizing the use of response times for item selection in CAT, J. Educ. Behav. Stat., advanced online publication (2017).
[51] D. Dunning, K. Johnson, J. Ehrlinger, and J. Kruger, Why people fail to recognize their own incompetence, Curr. Dir. Psychol. Sci. **12**, 83 (2003).
[52] J. C. Ban, B. A. Hanson, T. Wang, Q. Yi, and D. J. Harris, A comparative study of online pretest item-calibration/scaling methods in computerized adaptive testing, J. Educ. Measure. **38**, 191 (2001).
[53] J. C. Ban, B. A. Hanson, Q. Yi, and D. J. Harris, Data sparseness and on-line pretest item calibration-scaling methods in CAT, J. Educ. Measure. **39**, 207 (2002).
[54] H. Wainer and R. J. Mislevy, Item response theory, item calibration, and proficiency estimation, in *Computerized Adaptive Testing: A Primer*, edited by H. Wainer (Erlbaum, Hillsdale, NJ, 1990), p. 65.
[55] M. K. Hartwig and J. Dunlosky, Study strategies of college students: Are self-testing and scheduling related to achievement?, Psychon. Bull. Rev. **19**, 126 (2012).
[56] R. N. Blasiman, J. Dunlosky, and K. A. Rawson, The what, how much, and when of study strategies: Comparing intended versus actual study behavior, Memory **25**, 784 (2017).