Chained computerized adaptive testing for the Force Concept Inventory

Jun-ichiro Yasuda◉*

Center for the Studies of Higher Education, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan

Michael M. Hull◉

University of Alaska Fairbanks, Department of Physics, 1930 Yukon Dr, Fairbanks, Alaska 99775, USA

Naohiro Mae◉

Research Center for Nuclear Physics, The University of Osaka, Ibaraki, Osaka 567-0047, Japan

Kentaro Kojima◉

Kyushu University, Faculty of Arts and Science, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan

Although conceptual assessment tests are commonly administered at the beginning and end of a semester, this pre-post approach has inherent limitations. Specifically, education researchers and instructors have limited ability to observe the progression of students' conceptual understanding throughout the course. Furthermore, instructors are limited in the usefulness of the feedback they can give to the students involved. To address these challenges, we propose an alternative approach that leverages computerized adaptive testing (CAT) and increasing the frequency of CAT-based assessments during the course, while reducing the test length per administration, thus keeping or decreasing the total number of test items administered throughout the course. The feasibility of this idea depends on how far the test length per administration can be reduced without compromising test accuracy or precision. Specifically, the overall test length is desired to be shorter than when the full assessment is administered as a pretest and a subsequent post-test. To achieve this goal, we developed a CAT algorithm called *Chain-CAT*. This algorithm sequentially links the results of each CAT administration using collateral information. We developed the Chain-CAT algorithm using the items of the Force Concept Inventory (FCI) and analyzed its efficiency through numerical simulations. We found that collateral information significantly improved the test efficiency, and the overall test length could be shorter than that of the pre-post method. Without constraints for item balancing and exposure control, simulation results indicated that the efficiency of Chain-CAT is comparable to that of the pre-post method even when each CAT administration includes only 5 items and the CAT is administered 9 times throughout the semester. However, when these constraints are imposed, we found that the efficiency of the Chain-CAT with a test length of five items is lower than that of the pre-post method. Based on this result and analysis of which items were administered, we suggest expanding the FCI item bank by creating additional items or incorporating items from other research-based assessments that have high discrimination parameters and exhibit sufficient variability in their difficulty parameters.

I. INTRODUCTION

When measuring pedagogical effectiveness, it is common practice to administer assessment tests before and after instruction. After collecting the pre- and post-test scores, we calculate a statistic such as the average normalized gain and analyze the average change in students' understanding due to instruction. The results can then be compared with records from previous years or reference values in the literature to determine the effectiveness of the instruction. For example, the Force Concept Inventory (FCI) [1] is an assessment test used to probe students' conceptual understanding of Newtonian mechanics. The test has 30 items with five choices, and students typically take 20 to 30 min to complete it. Hake [2] compared the

---

average normalized gain of the FCI between interactive engagement courses and traditional courses and found that the interactive engagement method is more effective. There are many other recent studies that administered the FCI or other assessment tests both before and after instruction and analyzing the effects of that instruction in terms of improvement in students' scores [3–10].

The pre-post method is a standard approach to analyze the effectiveness of instruction, but there are limitations to the approach. First, the pre- and post-test results provide only snapshots of students' understanding at the beginning and end of the course, limiting the ability to observe the progression of conceptual understanding throughout the course. Second, students may see little benefit in taking the assessment when the focus of the assessment is to reflect on the year's instruction and improve instructional practices for future students, with no feedback to the students who actually take the assessment. It is therefore not surprising that post-test response rates are generally lower than pretest response rates, especially when students are asked to take the assessment outside of class, unless extra credit is offered to maintain student engagement and motivation [11–13].

To address these issues associated with the pre-post method, we propose an alternative approach that increases the frequency of assessments during the course while decreasing the test length per administration, thus maintaining or decreasing the total number of test items throughout the course. This idea is inspired by the microgenetic method [14–21], which involves frequent quantitative and/or qualitative measurements. The microgenetic method involves (i) sampling data at a frequency that is assumed to be high compared to the rate of change of the phenomenon of interest and (ii) collecting data for the entire period of change [20], allowing teachers to monitor the progression of students' understanding. Although these methods are generally implemented with small student cohorts, Sayre and Heckler [16,17] used the microgenetic approach by frequently administering simple multiple-choice conceptual questions to a large number of students.

While the pre-post method aligns with usage of diagnostic and summative assessments, our idea of frequent administrations of short tests can be used as a form of formative assessment. In terms of research instruments, the method of survey administration should be chosen depending on the research questions being investigated. In terms of enhancing instruction, our method opens the door to providing automated feedback based on each student's set of responses. We expect that this will increase the usefulness of the survey for students and reduce their sense of burden in completing it.

The feasibility of the approach we are proposing depends on how far the test length per administration can be reduced. A reference value of the total test length during a course is 60 items, as this is the total test length if the full FCI is administered both as the pre- and post-test. If the course involves ten administrations of the shortened FCI, it is then desired that the test length per administration would be less than six items.

To reach this goal, we use computerized adaptive testing (CAT) [22,23]. CAT is a practice in which a computer administers successive test items to match the current estimate of the student's proficiency (see below for details). In one popular model of CAT, if a student answers an item correctly, the student will next need to answer a more difficult item. On the other hand, if a student answers an item incorrectly, the student will then answer an easier item next. In this way, high (low) proficiency students do not need to answer items that are too easy (difficult) for them; thus the test length can be significantly shortened compared to standard test administration without compromising the accuracy and precision of the test. (Accuracy is the level of agreement between a measured value and a true value, and precision is the level of agreement between measured values obtained by replicate measurements on similar objects under specified conditions [24].) Because of its efficiency, CAT is becoming widely used, for example, with the Graduate Record Exam (GRE) [25], with the SAT [26], with PISA [27], and recently in physics and science education research [28–33].

We utilize a CAT algorithm that takes advantage of collateral information [22,34]. Collateral information refers to relevant empirical data about respondents, such as age, grade level, or previous test scores. In our study, we specifically used only the proficiency estimate from each respondent's previous test—that is, the most recent proficiency estimate carried over, as is commonly done (e.g., in Ref. [22]). This information can be used to select the first item in CAT, accelerating the convergence of the proficiency estimation, and hence improving test efficiency [22]. Specifically, as we describe below, we use the proficiency estimate of each respondent in a test for selecting items and estimating the respondent's proficiency level in the next test for the respondent. In contrast to one-time testing, these subsequent tests constitute multiple measurements over time (e.g., with intervals of about a week between them). Since this CAT algorithm sequentially links the test result of each administration, we call this algorithm *Chain-CAT*. In CAT, the proficiency estimate is updated after each item and used to select the next item at a given point in time using a specified algorithm. In addition to this, Chain-CAT uses collateral information, employing the same algorithm with the proficiency estimate from the last question of the previous test (e.g., one week before) to initiate item selection in the current test. Although the algorithm is the same, the term "collateral information" is introduced to emphasize the time gap between tests, during which learning can occur.

Figure 1 illustrates the Chain-CAT algorithm in the case when a student takes the tests for a course with seven administrations of the Chain-CAT. Each test is
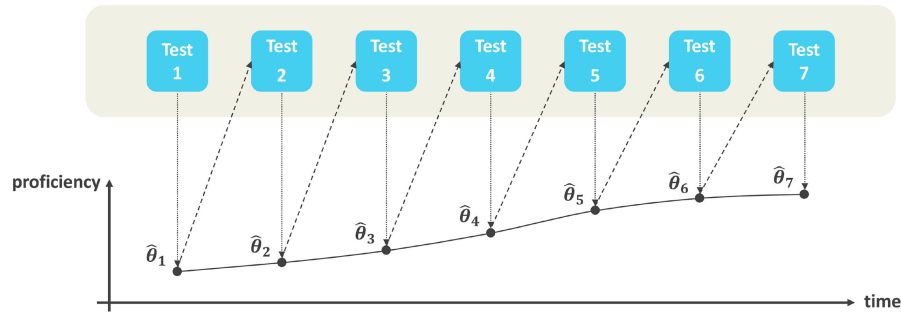
FIG. 1.   Illustration of Chain-CAT algorithm for a student taking the tests for a course with seven administrations of the Chain-CAT.

computerized adaptive, and each test length is relatively small (e.g., ranging from five to ten items). After the student takes test 1, his or her proficiency level is estimated (represented as $\hat{\theta}_1$ in the figure), and it is used as collateral information for the item selection and proficiency estimation at test 2. This procedure is continued sequentially from the next test to the final test (test 7).

Several researchers have studied the development of a CAT-based formative assessment system. Veldkamp *et al.* [35] discussed the viability of a computerized adaptive learning system which responds to students based on their test results. Choi and McClenen [36] developed and analyzed the validity and efficiency of an adaptive formative assessment system based on CAT with the time series proficiency estimation utilizing Bayesian networks. Yang *et al.* [37] developed a formative assessment system based on CAT and the learning memory cycle. While there have been previous studies such as these, our goal is to construct an algorithm from our design of chaining collateral information as described above and to introduce its practical application in the context of physics education research.

The objectives of Heckler and Sayre [17] and our Chain-CAT are partially aligned, particularly in terms of tracking student understanding. However, the methods employed differ significantly. Heckler and Sayre [17] repeatedly administered individual questions to randomly selected groups of students and analyzed changes in the correct response rate of each question. In contrast, our approach employs a CAT-based system in which the test consists of multiple questions. The selection of those questions varies depending on the student's responses, and the analysis is based on the proficiency estimates derived from those responses.

The objective of this study is to examine the feasibility of the Chain-CAT version of the FCI (Chain-CAT FCI). Specifically, our research questions are: (i) How much does collateral information improve the test efficiency when it is used to sequentially link the results of each CAT administration?, (ii) Is it possible to shorten the total test length of Chain-CAT FCI in a course to be less than that of the full FCI conducted as the pre- and post-test (60 items) without compromising test accuracy and precision, and under what conditions is this possible?, and (iii) Do the

30 FCI items comprise a sufficient item bank for the effective use of Chain-CAT? If not, what improvement is necessary? The focus of this paper is on the analysis of the feasibility of the Chain-CAT FCI using numerical simulations. The development of feedback methods to deliver formative assessment will be the subject of future work.

The remainder of this paper is organized as follows: In Sec. II, we present the design of the algorithm for Chain-CAT and the numerical simulation procedure for analyzing its efficiency. In Sec. III, we describe the simulation results and discuss the feasibility of the Chain-CAT FCI. Finally, in Sec. IV, we summarize this study and discuss its limitations and future work.

All our analyses were conducted using R [38]. In addition to the basic package of R, the item parameters of the FCI were calibrated using the package MIRT [39], and the simulations of the Chain-CAT FCI were conducted using the package catR [40,41].

II. METHODOLOGY

A. Item response model and item bank

We constructed the item bank of the Chain-CAT using the 30 FCI items. The item parameters were calibrated based on the two-parameter logistic (2PL) model [42]. In this model, the probability of a correct response from the ith respondent on item j is given by

$$P_j(\theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \qquad (1)$$

In Eq. (1), θ_i is the parameter representing the proficiency of the ith respondent. The proficiency distribution in a reference population is standardized; namely, the estimated mean of θ_i is set to 0 and the estimated standard deviation of θ_i is set to 1. In the equation, b_j is the difficulty parameter, and a_j is the discrimination parameter of item j. The items with higher a_j can better distinguish respondents who have different levels of proficiency.

We estimated the item parameters using the FCI survey data obtained in our previous study [43,44]. In that study, we administered the full-length paper-and-pencil FCI in

class to 2882 Japanese university students from April 2015 to April 2018. The examinees were students at the beginning of introductory physics courses at one public and four private universities, all of which are considered mid-ranked institutions in Japan. From this dataset, we removed aberrant responses, leaving 2712 valid responses. The result of the item parameter estimation is shown in Table III in the Appendix. Although the standard error of the difficulty parameter for question 29 is relatively large, this item was rarely administered in the simulation due to its low discrimination. Hence, we expect that its inclusion has no significant effect on our results.

The assumptions of unidimensionality, overall local independence, and goodness of fit were examined in our previous studies [43,44], supporting the validity of using the 2PL model in our analysis. Specifically, we assessed the unidimensionality of the FCI using principal component analysis with a tetrachoric correlation matrix [42] and evaluated the assumption of local independence using Yen's Q_3 statistic [45]. The results indicated that our FCI dataset has sufficient unidimensionality and local independence at the overall test level, although higher unidimensionality and local independence are desirable (see Sec. IV B).

We evaluated the goodness of fit of 1PL, 2PL, and 3PL IRT models to the response data using the standardized root mean square residual (SRMSR) [46]. A value of SRMSR less than 0.05 is considered to indicate that the model is well fitted. The SRMSR values obtained were 0.079 for the 1PL model, 0.041 for the 2PL model, and 0.041 for the 3PL model. These results suggest that both the 2PL and 3PL models fit the data well, whereas the 1PL model does not.

To select the most appropriate model, we used the Bayesian information criterion (BIC) [47]. BIC increases if the deviance of the model from the data increases and if the number of parameters increases; thus, the model with the lowest BIC is the most preferred. Based on our dataset, the BIC was 9.22×10^4 for the 1PL model, 9.00×10^4 for the 2PL model, and 9.00×10^4 for the 3PL model. These results indicate that the BIC values for the 2PL and 3PL models are comparable and both are lower than that of the 1PL model.

Given the possibility that future studies may compare their results with those of the present study, it is important to reduce the cost of data collection. Since the recommended sample size for accurately estimating item parameters is approximately 500 for the 2PL model and around 2000 for the 3PL model [42], we therefore selected the 2PL model for our analysis.

B. Basic computerized adaptive testing process

The CAT process consists of four successive steps [23]: (i) initial step, (ii) test step, (iii) stopping step, and (iv) final step. The basics for each step are as follows.

(i) *Initial step*: The first item is selected and administered to a respondent. The most commonly used criterion to select the first item is the maximum Fisher information (MFI) criterion [23]. The MFI criterion calls for selecting the most informative item (the item with the largest Fisher information) for the respondent based on the current estimate of proficiency. (Generally speaking, the "most informative item" is the one that will minimize the standard error of the proficiency estimate [42].) When nothing is known about the respondent, as is often the case when the first item is chosen for the first test, the Fisher information of the item is calculated using the mean proficiency value of the prior population, which is most often set to zero.

(ii) *Test step*: The proficiency of the respondent is estimated using the current set of item responses, and the next item is selected to be administered. For the proficiency estimation method, as in our previous study [43,44], we used the expected *a posteriori* (EAP) method. For the item selection criterion, we used the MFI criterion as in the initial step.

(iii) *Stopping step*: The test checks that a certain criterion has been met, and the administration of the items ends. We chose test length to be the stopping criterion, such that CAT stops after a predetermined number of items have been administered, ranging from 1 to 30 items taken from the FCI.

(iv) *Final step*: The final step involves the calculation of the final estimate of the respondent's proficiency level. As in the test step, we chose the EAP method to estimate the proficiency level.

C. Design of the Chain-CAT algorithm

In the Chain-CAT algorithm, as we described above, we linked the results of each CAT administration sequentially using collateral information. Specifically, each test incorporates the proficiency estimates of previous tests as collateral information in the prior distribution for the proficiency estimation based on the Bayesian method. This generally accelerates the convergence of the estimates in each CAT administration, thereby improving test efficiency.

There are three stages where we can utilize collateral information: the initial step, the test step, and the final step of the testing process. This process is completed within each "test" in Fig. 1. In what follows, we explain how we implemented collateral information for each stage.

(1) *Initial step*: In this step, the first item is selected and administered to a respondent, as we described above. For example, when the MFI criterion is used, the algorithm selects the item with the largest Fisher information for the current estimate of the respondent's proficiency. At the beginning of the first test, when we know nothing about the respondents, the Fisher information of the candidate items is calculated using the mean proficiency value of the prior

population. This value is commonly set to zero to have the scale centered on respondents [23], as we described above.

At the beginning of each test after the first one, in the Chain-CAT algorithm, the Fisher information can be calculated utilizing the proficiency estimate of the former test for a given student as collateral information to improve the test efficiency for that student. More generally, at the beginning of the nth ($n \geq 2$) test, the Fisher information can be calculated utilizing the $(n-1)$th test proficiency estimate of a given student as collateral information.

(2) *Test step*: At this stage, CAT selects the next item based on the estimated proficiency level of the respondent at that point in the test. We estimated the proficiency level using the EAP estimator, which is based on the Bayesian posterior distribution. The posterior distribution for an ith respondent in turn is proportional to the product of the likelihood function and a prior distribution of the proficiency $g(\theta^i)$ [23]. We used a model with a normal distribution; thus, we represent $g(\theta^i) \sim \mathcal{N}(\mu^i, \sigma^i)$, with a mean of μ^i and a standard deviation of σ^i. (Assuming a normal distribution is common in proficiency estimation using Bayesian item response modeling [23].)

On the first test, when we know nothing about the respondents beforehand, a common choice of the prior distribution is a normal distribution with $\mu^i = 0$. On the second test, we can utilize the first test proficiency estimate of an ith respondent, $\hat{\theta}^i_1$, as collateral information for the prior distribution. Specifically, we chose the prior distribution as a normal distribution with $\mu^i = \hat{\theta}^i_1$ for the second test. To generalize, we used the $(n-1)$th test proficiency estimate of an ith respondent, $\hat{\theta}^i_{n-1}$, as collateral information to determine the normal distribution such that $\mu^i = \hat{\theta}^i_{n-1}$ in the prior distribution for the nth test.

The standard deviation of the prior distribution cannot be estimated in the method directly using the proficiency estimate. In the simulation analysis (described in more detail below), we analyzed how the test efficiency depends on the value of σ^i and searched for the optimal value which maximizes the accuracy and precision.

(3) *Final step*: We used collateral information also for the final proficiency estimation using the EAP method, as just described for the test step.

D. Improvement of the validity of Chain-CAT

To increase the validity of the Chain-CAT FCI as a test of student understanding of Newtonian mechanics, one can impose constraints on how items are selected [23]. First, we implemented *content balancing* for the Chain-CAT FCI as

in our previous study [48]. Content balancing ensures that the items in each test are not biased toward specific concepts (e.g., Newton's third law), and the same set of concepts assessed in the original test is covered in the CAT administration for each respondent. To balance content in CAT, the percentage of items to be administered from each subgroup is defined in advance, for example, to be the same as that found in the FCI itself. Doing so ensures that items from each subgroup are administered. The number of items and percentage of total items in each subgroup on the FCI are [48]: Kinematics (5 items, 17%), first law (8 items, 27%), second law (3 items, 10%), third law (4 items, 13%), and kinds of forces (10 items, 33%).

Various algorithms exist to control content balancing, but the catR package [23] allows only the use of the simplest option, the constrained content balancing method [49]. The content balancing algorithm begins with the second item administered. The steps of the algorithm for selecting the second and subsequent items are [48]:

1. Before the administration of each item, compute the percentage of items that have already been administered from each subgroup.

2. Target the optimal subgroup of items. Generally, this is the subgroup for which the gap between the observed relative proportion of administered items and the expected relative proportion is maximal. When multiple subgroups have not yet had any of their items administered, then one of these subgroups is chosen at random.

3. Once the subgroup for the next item has been chosen in step 2, an item is selected from this subgroup to be the next item administered. As described above, the MFI criterion is used to choose the item from within the subgroup.

For example, suppose that question 13 is selected for the first item. This item is from the kinds of forces subgroup. Following the algorithm above, the second item will come from a subgroup that is chosen randomly from the remaining four subgroups. After an item from each subgroup is administered, the sixth item will be chosen from the kinds of forces subgroup, since the gap between the observed relative proportion (20%) and the expected relative proportion (33%) is maximal.

Second, we controlled the *item exposure* for the Chain-CAT FCI. Generally, if the items in an item bank are very informative at the average ability level, they are selected more frequently than other items [23]. Although a given item does not appear more than once in a given test, the simple Chain-CAT algorithm can result in that item appearing frequently across the tests. If students repeatedly answer certain items each week, they may memorize the content of the items they solved the previous week and answer them without rethinking, which may not reflect their actual proficiency. Therefore, we implemented an algorithm that limits the number of times an item can be

administered in a course. Specifically, we used an algorithm whereby once an item is administered a specific number of times in a course (ranging from three to nine items, depending on the test length), it is removed from the list of items that can be administered (described in more detail below).

E. Numerical simulation procedure

We conducted a numerical simulation to analyze the efficiency of the Chain-CAT and to search for the optimal settings. The outline of the procedure is as follows. (i) Assume a progression model of the true-value proficiency θ as a function of session s (the unit is taken to be 1–2 weeks) in a course. (ii) Generate response data for a given true value of θ based on the item response model and calculate the estimates of θ. (iii) Calculate the accuracy and precision by the root-mean-square error (RMSE) from the true and the estimated values of θ and compare it to a reference value. In what follows, we explain the details of the above procedure.

1. Progression models of the true proficiency level

To analyze the efficiency of the Chain-CAT as generally as possible, we assumed various progression patterns of the true-value proficiency θ. Specifically, we assumed the following progression models for $\theta(s)$: (a) a stationary model in which θ is constant across sessions, (b) a linear model in which θ increases linearly across sessions, and (c) a step model in which θ increases significantly only at a certain session. Although it is uncommon in introductory physics courses for student scores on the FCI to decrease, we confirmed that the simulation results of the case when θ decreases linearly or decreases significantly only at a certain point are similar to the results of (b) or (c), respectively. We supposed that most of the student's progressions are one of these patterns or a combination of them. Moreover, we analyzed the test efficiency of the Chain-CAT in various proficiency ranges for each model. For example, in the stationary model, we analyzed the efficiency in the middle θ students ($\theta = 0$) and the high θ students ($\theta = 1.0$). We confirmed that the low θ range ($\theta = -1.0$) yields similar results to the high θ range ($\theta = 1.0$), so we do not show the result below. We fixed the number of tests conducted throughout the simulated course to be 9 times. This would be a reasonable number of tests to be administered in a course consisting of 10–15 weeks.

2. Generation of the response data

We generated response data based upon a Monte Carlo simulation, which is commonly used in the development of CAT [50]. Specifically, we generated the responses for Chain-CAT using the function *simulateRespondents* of the catR package. In the simulation, the response to a selected item is generated based on the calibrated item parameters of that item and the value of true proficiency at a particular session. For example, in Eq. (1), suppose that the probability of a correct response for an item is calculated to be 0.75 for a given true proficiency. A random number is generated from a uniform distribution within a range of 0 to 1. If the value is 0.75 or less, the generated response is coded as "correct" for that item. If the value is greater than 0.75, then the generated response is "incorrect." The EAP method is used to estimate the proficiency of the respondent, an item selection method (e.g., MFI criterion) is then used to choose the next item based upon that estimated proficiency, and the process repeats until reaching the predetermined test length. Specifically, we varied the test length to be $L = 5, 10,$ and 15. Then, for each length, we calculated accuracy and precision.

Note that when the test is administered 9 times with test lengths of $L = 5, 10,$ and 15, the total number of items becomes 45, 90, and 135, respectively. In contrast, the pre-post method uses a total of 60 items. Therefore, only in the case of $L = 5$ is the total number of items fewer than in the pre-post method. Our analysis in this paper includes results for $L = 10$. We think such test lengths may be viable for instructors, even if the total number of questions exceeds 60. Limiting each session to ten questions helps distribute the workload, potentially reducing the perceived burden. Moreover, even if the total number of questions is greater, having multiple small tests throughout the semester allows instructors to provide feedback as part of formative assessment.

To summarize, provided calibrated item parameters and a given true proficiency level, the *simulateRespondents* function generates an entire set of responses that are either correct or incorrect and calculates estimated proficiencies for a given test length of each Chain-CAT administration. This is done for each session in a course. In this analysis, we generated 10 000 response data based on the same true theta for a given session.

3. Algorithm to improve the validity of Chain-CAT

As we mentioned above, to improve the validity of Chain-CAT, we balanced the content and reduced the exposure of individual items. In the constrained content balancing method, items are selected so that the proportion of concept subgroups in an administered test is similar to that of the original test. Let us assume the case of $L = 5$, for simplicity. The FCI has five subgroups ($N_{\text{sub}} = 5$), as discussed above. Thus, $L/N_{\text{sub}} = 1$ item from each subgroup is administered in each test. If we administer the test at N_S (=9 in our simulation) sessions throughout the duration of the course, the total number of items to administer from each subgroup, N_{tot}, is thus given by N_S times L/N_{sub}, which equals 9.

In addition, to reduce exposure of individual items, we used an algorithm whereby once an item is administered a specific number of times, $N_{\text{adm}}(L)$, to a given respondent in

a course, it is removed from the list of items that can be administered to that respondent. The threshold N_{adm} should be as small as possible; however, since there are only a limited number of items in the FCI and each subgroup, there is a lower limit to N_{adm}. If the number of times to administer a given item is set too low, there will be no items left to administer before the course has concluded.

For example, the second law subgroup has three items ($N_{\text{cb}} = 3$). If $L = 5$ and $N_{\text{S}} = 9$, the total number of items to administer from each subgroup, N_{tot}, is equal to 9 from the above calculation, so each item in the subgroup should not be removed until it is administered 3 times ($= N_{\text{tot}}$ divided by N_{cb}). Therefore, in this case, we determine N_{adm} to be 3. For consistency, we use the same threshold in the other subgroups as well: for $L = 5$ and $N_{\text{S}} = 9$, a given FCI item is not administered more than 3 times.

For the cases where the test length is longer than five items, the calculation of N_{adm} is more complicated because the frequency of administering items is no longer the same for each subgroup. We determined N_{adm} by manually creating a table to lay out the order in which subgroups would be chosen for each subsequent item. We then counted the number of items administered from each subgroup and derived that $N_{\text{adm}}(10) = 5$, and $N_{\text{adm}}(15) = 9$ are the minimum number of times administered of a given item without resulting in subgroups becoming unrepresented in tests administered later in the semester. We confirmed these thresholds using numerical simulations. Note that for $L = 15$, $N_{\text{adm}} = 9$ means that there is no limit because it equals the number of tests, $N_{\text{S}} = 9$. That is, for $L = 15$, a given FCI item can appear on each of the nine tests.

4. Calculation of the accuracy and precision

We represented the accuracy and precision by the bias and variance, respectively. Then, we summarized these measures in terms of the mean-square error (MSE). The MSE is a commonly used statistic for evaluating the accuracy and precision of parameter estimation in simulation studies. It is effective when the true proficiency parameters (theta) are known, particularly in the context of CAT [23]. The MSE is defined by the following equation [51], which equals the sum of the variance and the squared bias:

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var} + B^2, \qquad (2)$$

where $E(x)$ is the expected value of x. In the equation, θ is the true proficiency given for each session and $\hat{\theta}$ is the estimated proficiency level. The MSE was calculated as the average of the squared difference of θ and $\hat{\theta}$ using the above 10 000 response data for each session. The bias B is defined by

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta, \qquad (3)$$

where $E(\hat{\theta})$ is represented by the average of the estimated proficiency level. The variance is given by

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E_{\hat{\theta}})^2], \qquad (4)$$

where $E_{\hat{\theta}} = E(\hat{\theta})$. With the above definitions, we show our simulation results with the square root of the above statistics, namely, the root-mean-square error (RMSE) defined by RMSE $= \sqrt{\text{MSE}}$ and the standard deviation SD $= \sqrt{\text{Var}}$.

Our previous study [43,44] indicated that, within the 30 items of the FCI, longer test lengths led to smaller RMSE. Although generally, the higher the accuracy, the better, our approach prioritizes features such as test time and application to formative assessment, and so we look for a degree of accuracy that is comparable to that of the paper-based pre-post method. Specifically, we calculated the RMSE for the Chain-CAT and compared it to the RMSE obtained using the conventional pre-post method with the FCI. In the latter approach, all 30 FCI items were administered in both the pretest and post-test, and student proficiency was estimated using item response theory based on the same 2PL model described above, without collateral information. The RMSE from the pre-post method served as a reference value for evaluating the efficiency of the Chain-CAT.

When comparing the accuracy and precision of the Chain-CAT with the pre-post method, we use two different types of averages for the RMSE. The first one is the session-averaged RMSE, which is defined by

$$\text{session.av.} \, \text{RMSE} = \frac{\sum_{s=s_i}^{s_f} \text{RMSE}(s)}{N_{\text{S}}}, \qquad (5)$$

where s_i is the initial session, s_f is the final session, and N_{S} is the number of the sessions (number of tests, as above). The second type of average RMSE we calculated is the overall average of the RMSE, which is calculated by taking the average of the session-averaged RMSE for all progression models that we assumed in the simulations:

$$\text{all.av.} \, \text{RMSE} = \frac{\sum_{m=1}^{N_{\text{M}}} \text{session.av.} \, \text{RMSE}(m)}{N_{\text{M}}}, \qquad (6)$$

where m is the index of the progression model and N_{M} is the number of the progression models we considered.

III. RESULT

A. Typical numerical simulation result

Figure 2 shows the simulation results assuming a linear model for the true proficiency progression ($0 \leq \theta \leq 1$). The test length of each administration was set to $L = 10$. The standard deviation of the prior distribution was fixed at $\sigma^i = 0.5$. The left figure of Fig. 2 is a result without collateral information (basic CAT), and the right figure is a result with collateral information. The graphs include the simulation results for nine sessions from s_1 to s_9, where the true proficiency level increases from $\theta(s_1) = 0$ to $\theta(s_9) = 1.0$ (see the purple lines in the graphs). For each session, we
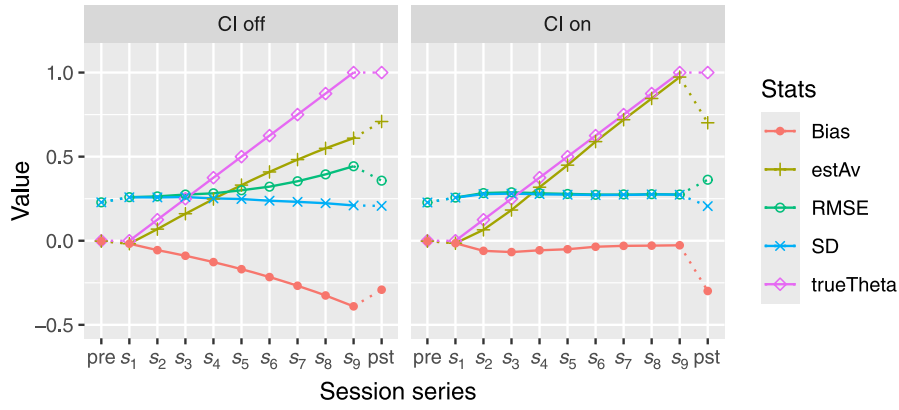
FIG. 2.   Typical numerical simulation results for a linear progression model. The test length of each administration is set to $L = 10$. The standard deviation of the prior distribution was fixed at $\sigma^i = 0.5$. Left: without collateral information (basic CAT). Right: with collateral information (Chain-CAT). The values of "pre" and "pst" show the results based on the pre-post method where the full FCI (30 items) is used in the proficiency estimation without collateral information.

generated 10 000 responses based on the same true theta at a given session and estimated the proficiency level for each response, then we calculated the average and the standard deviation of the estimated theta (the yellow and blue lines). The standard deviation for the estimates of the θ mean is about 0.0025. The bias was calculated from the difference between the true theta and the average of the estimated theta (the red lines), then the RMSE was calculated using Eq. (2) (the green lines). In the right figure of Fig. 2, the standard deviation is much larger than the bias, so the SD values and the RMSE values are very close. In the graphs, the reference values are shown at $s = $ pre and $s = $ pst on the horizontal axis, where the statistics were calculated using the true proficiency levels at $s = s_1$ and $s = s_9$ and administering the whole test items (30 FCI items) in both cases without collateral information.

From Fig. 2, we can observe that the collateral information significantly improved the accuracy. With collateral information, the Fisher information is calculated using the proficiency level of the previous test, which prevents the bias from becoming larger when the true proficiency level is away from $\theta = 0$. Without the collateral information, the Fisher information is always calculated assuming the true proficiency level equals zero, therefore, the deviation between the true theta and the estimated theta becomes large when the true proficiency level is away from $\theta = 0$. Even if we use the 30 FCI items, the bias is quite large when $\theta = 1.0$, as shown at $s = $ pst in Fig. 2.

B. Comparing overall trends

As we mentioned above, the standard deviation of the prior distribution σ^i is a free parameter. We found that if σ^i is small, the variance in Eq. (4) tends to be small, but the bias in Eq. (3) tends to be large when the proficiency progression is large. If σ^i is large, the variance tends to be large, but the bias tends to be relatively small when the

proficiency progression is large. Moreover, we found that the rank order of the overall average RMSE magnitude depends on the value of σ^i. For example, when σ^i is set to 1 for all methods, the pre-post method shows the highest accuracy (lowest RMSE), whereas when σ^i is set to 0.3, the Chain-CAT with collateral information shows the highest accuracy, and so on. We were concerned that comparing results at a fixed σ^i could introduce arbitrariness, depending on the analyst's choice of the common σ^i value. Therefore, we calculated the overall average across various σ^i values and based our comparison on the σ^i that yielded the lowest overall average RMSE. This approach reflects practical considerations for implementing the test in classroom settings, where σ^i should be determined before administration without knowing the amount of proficiency progression in advance. In this case, it is reasonable to choose σ^i that yielded the lowest overall average in each method.

Specifically, for a given value of $\sigma^i (= 0.1, 0.2, \ldots, 1.0)$, we conducted simulations and calculated the session-averaged RMSE in Eq. (5) from $s = s_1$ to $s = s_9$. For the pre-post method, the session-averaged RMSE was calculated for $s = $ pre and $s = $ pst. Then we calculated the overall average in Eq. (6) for all progression models that we assumed in the simulations. After that, we chose the value of σ^i which minimizes the overall average of the RMSE. We obtained the following values for each case, namely, $\sigma^i = 0.4$ for the case without collateral information, $\sigma^i = 0.3$ for the case with collateral information, and $\sigma^i = 0.6$ for the pre-post method.

After fixing the value of σ^i which is most likely to give the highest efficiency for each case, we calculated the session-averaged RMSE. Table I illustrates the effect of collateral information within each progression model for the case of $L = 10$. For each model, we compare the effectiveness of different methods: without CI, with CI, and the pre-post approach. To isolate and better understand the

TABLE I. The session-averaged RMSE for each progression model ($L = 10$).

Progression model	CI off	CI on	Pre-post
Constant ($\theta = 0$)	0.210	0.186	0.252
Constant ($\theta = 1$)	0.531	0.289	0.322
Linear ($0 \le \theta \le 1$)	0.332	0.233	0.287
Linear ($-0.5 \le \theta \le 0.5$)	0.248	0.248	0.260
Linear ($-1.0 \le \theta \le 0$)	0.347	0.322	0.293
Step ($0 \le \theta \le 0.5$)	0.262	0.214	0.256
Step ($-0.5 \le \theta \le 0$)	0.250	0.250	0.256
All model average	0.311	0.249	0.275

impact of collateral information, we did not incorporate content balancing or item exposure control in this table. In Table I, we can observe that the session-averaged RMSE in the case with collateral information is smaller than or comparable to that in the case without collateral information for all progression models. Moreover, the session-averaged RMSE in the case with collateral information is smaller than that of the pre-post case for most of the progression models. Especially, the case with collateral information is much more efficient when the absolute value of θ at final s is large as in the constant model ($\theta = 1$) and the linear model ($0 \le \theta \le 1$). On the other hand, the case without collateral information has a comparable efficiency in the linear model ($-0.5 \le \theta \le 0.5$) and the step model ($-0.5 \le \theta \le 0$), where the absolute value of θ varies near zero and the absolute value of θ at final s is small. One of the causes of this could be that the case without collateral information always assumes the mean value of prior distribution $\mu^i = 0$. Finally, the pre-post case is more efficient when the absolute value of θ at initial s is large and at final s is close to 0 as in the linear model ($-1.0 \le \theta \le 0$). This is because for CAT with short test length, such as $L = 10$, if the absolute value of θ at $s = s_1$ is large, the RMSE at $s = s_1$ is much larger than that of the pretest where $L = 30$.

C. Length dependence, content balancing, and item exposure control

Table II shows the dependence of the overall average RMSE of Chain-CAT (with collateral information) on the

TABLE II. The overall average RMSE for each test length with and without content balancing and item exposure control. The overall average RMSE is calculated by taking the average of the session-averaged RMSEs across all progression models assumed in the simulations, as defined in Eq. (6). These results are for the case with collateral information.

Test lengh	5	10	15
CB and IE off	0.277	0.249	0.233
CB and IE on	0.317	0.269	0.233

test length of each administration. The overall average RMSE is calculated by taking the average of the session-averaged RMSEs across all progression models assumed in the simulations, as defined in Eq. (6). The standard deviation of the prior distribution was fixed at the optimal value, $\sigma^i = 0.3$. Table II also includes a comparison of the cases with and without content balancing and item exposure control. In Table II, we can see that the overall average RMSE increases as the test length decreases. However, the overall averaged RMSE of Chain-CAT is comparable to that of the pre-post method even when the test length is $L = 5$ (compare the value of 0.275 in Table I to that of 0.277 in Table II). This means that if we administer the Chain-CAT 9 times in a course, by using the collateral information, we could reduce the total test length in a course to 45 items from the pre-post method (60 items) without compromising test accuracy and precision if we are not concerned about improving the validity of the test. On the other hand, by implementing content balancing and item exposure control, the overall average RMSE is increased from 0.277 to 0.317 when $L = 5$, an increase of about 10%. This is because when the choice of administration is limited, the items that result in lower accuracy and precision are administered. When $L = 10$, the overall average RMSE with content balancing and item exposure control is smaller than that of the pre-post method (compare the value of 0.275 in Table I to the value of 0.269 in Table II). In this case, the total test length is 90 items which is longer than that of the pre-post method, but the accuracy and precision are increased. This improvement is partly attributable to the assessment using 50% more items.

As noted above, applying content balancing or item exposure control limits the flexibility of item selection, resulting in the administration of items that reduce accuracy and precision. The findings above indicate that the current 30-item FCI is insufficient for implementing Chain-CAT effectively, and that expanding the item bank—either by developing new items or incorporating items from other concept inventories—is necessary. In this context, the results shown in Tables I and II, obtained without content balancing or item exposure controls, offer valuable insights. They can be interpreted as representing a future scenario in which the FCI item bank has been expanded to include a sufficient number of replacement items with similar difficulty and discrimination characteristics. In other words, from these findings, we expect that such a level of measurement accuracy could be attainable if the FCI item bank is expanded, even with content balancing and item exposure controlled.

Figure 3 shows histograms of how many times an item was administered in the nine tests when $L = 5$ in the linear model ($0 \le \theta \le 1$). The horizontal axis shows the number of times an item was administered, and the vertical axis shows the frequency. If an item appears in all nine tests, its frequency contributes one count to the bin labeled "9" in
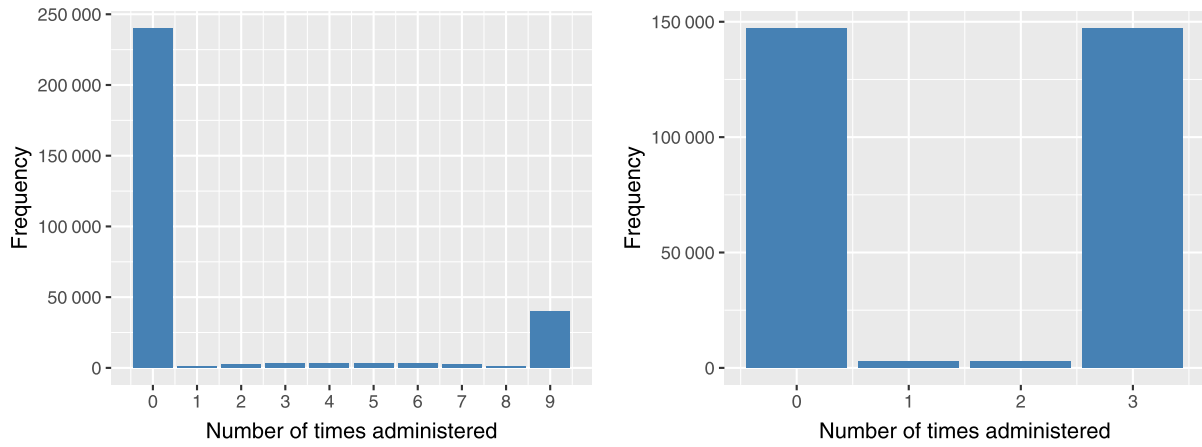
FIG. 3. Histograms of the number of times an item was administered in the nine tests when $L = 5$ in the linear model ($0 \leq \theta \leq 1$). If an item appears in all nine tests, its frequency contributes one count to the bin labeled "9" in the histogram representing the number of times administrated. The left (right) graph shows the result without (with) content balancing and item exposure control. The sum of the products of each x value (number of times administrated) and its corresponding y value (frequency) equals the total number of item administrations in the simulation, which is 450 000.
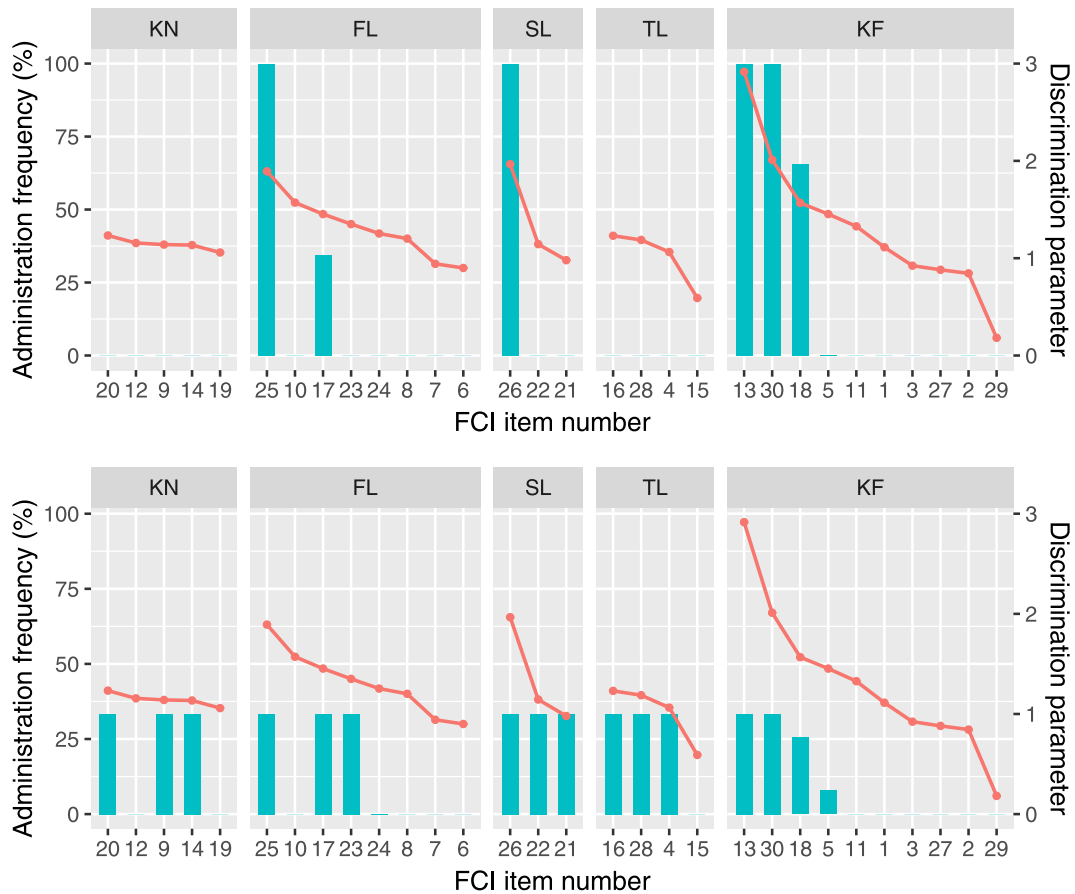

FIG. 4. Histogram of the likelihood of an item to be administered on a given test when there are nine tests of $L = 5$ in the linear model ($0 \leq \theta \leq 1$). The red lines show the discrimination parameters for the FCI items. The graphs are divided by subgroup of the FCI: Kinematics (KN), first law (FL), second law (SL), third law (TL), and kinds of forces (KF). The top (bottom) graph shows the result without (with) content balancing and item exposure control.

the histogram representing the number of times administrated. The left (right) graph shows the result without (with) content balancing and item exposure control. Recall that our simulation included 10 000 simulees taking nine tests. For $L = 5$, this corresponds to a total of 450 000 items administered in the simulation. In the histogram, the sum of the products of each x value (number of times administered) and its corresponding y value (frequency) equals the total number of item administrations in the simulation, which is 450 000. If we count any time that any of the 30 FCI items showed up in all nine of a simulee's tests, we find that occurs many times when we do not employ content balancing and item exposure control (rightmost bar of the left graph in Fig. 3). Conversely, there are also many cases in which an item is never administered (leftmost bar of the left graph in Fig. 3) during the semester.

Figure 4 (top) shows histograms of the likelihood for an item being administered in each test when there are nine tests of $L = 5$ in the linear model ($0 \leq \theta \leq 1$). The graphs are divided into subgroups on the FCI: Kinematics (KN), first law (FL), second law (SL), third law (TL), and kinds of forces (KF). We can see that some of the items with high discrimination parameters (Q13, Q25, Q26, and Q30) have a nearly 100% probability of being administered. That is, a student would see these four items on each of the nine tests they take. The fifth item we see would alternate between Q17 and Q18. The other 24 items were very rarely administered to the simulees.

The implementation of content balancing and item exposure control changes this situation, as shown in the right graph of Fig. 3. First, for content balancing, we ensured that the items in each test are not biased toward specific concepts. For example, since there are five subgroups, there was one item from each subgroup on each test when $L = 5$. In addition to this, following the calculation in Sec. II E 3, we set $N_{adm} = 3$ as the maximum number of times each item could be administered in the nine tests for $L = 5$. We can see that these are realized in the right graph in Fig. 3. However, even with content balancing and item exposure control, there are many cases where an item is never administered in that semester. This problem is evident in the bottom graph in Fig. 4. Compared to the top graph, the number of items with nonzero percentages has increased. However, nearly half of the items with low discrimination were never administered. (Although Q10 and Q12 have relatively high discrimination, they were almost never administered because their difficulty parameters are very small (less than -1.4). One additional issue concerns the variability in difficulty parameters. Since three of the content areas (KN, SL, and TL) contain five or fewer items, the range of difficulty parameters is limited. For example, the difficulty parameters of the SL items are 0.471 for Q21, -0.125 for Q22, and 0.567 for Q26. This limited variability may be insufficient to accurately and precisely estimate students' abilities.

These results show that simply expanding the item bank may be insufficient to improve the situation further. To improve the test efficiency, we suggest creating or incorporating items from other assessment tests [52–55] that are highly discriminative, i.e., that can better distinguish respondents with different levels of proficiency. In addition, the items should display sufficient variability in their difficulty parameters.

IV. DISCUSSION

A. Summary

To address the limitations of the pre-post method, we proposed an alternative approach using a Chain-CAT algorithm that sequentially links the results of each CAT administration in a course using collateral information. We analyzed the efficiency of the Chain-CAT algorithm through numerical simulations of various progression models for the values of true proficiency. The simulation results showed that the Chain-CAT algorithm significantly improved the test efficiency in most of the progression models compared to the simple CAT without collateral information. The simulation results indicated that the accuracy and precision of the Chain-CAT are comparable to that of the pre-post method even when each test length is as short as 5 items when the test is administered 9 times; therefore, the total test length of Chain-CAT can be less than that of the pre-post method. This finding, however, is without imposing constraints of item balancing and item exposure control. When these constraints are imposed, we found that the efficiency of the Chain-CAT is lower than that of the pre-post method when the length of each of the nine tests is five items. These results suggest that the item bank should be expanded by adding new items or incorporating items from other inventories that have high discrimination and exhibit sufficient variability in difficulty parameters.

B. Limitations and future work

1. Validating Chain-CAT further

There are many possibilities for further improving the validity of the Chain-CAT. First, although we found that our FCI dataset has sufficient unidimensionality and local independence at the overall test level [43,44], higher unidimensionality and local independence are desirable. This can be realized by dropping one item from each locally dependent pair. Adding new items would be necessary to compensate for the reduction in efficiency. Alternatively, using multidimensional models may produce a better fit to the data, resulting in more valid proficiency estimates [56–59].

Second, our content balancing algorithm utilized the FCI taxonomy table [48], but researchers have identified issues with this grouping system [54,60–65]. When new highly discriminative items (as mentioned above) are added to the

item bank, the reliability of the statistical grouping and the content balancing algorithm should be examined.

Third, it is meaningful to conduct Chain-CAT on real classes and assess the validity of the instrument. We can analyze the validity of the real class data by analyzing the consistency between the Chain-CAT results and other assessment data, such as that obtained in clinical interviews. We have conducted a pilot study for validation, where we deployed the Chain-CAT system using the mirtCAT package [66] with shinyapps. Students used their smartphones to complete the Chain-CAT. The preliminary results showed that Chain-CAT results are roughly consistent with interview findings. Moreover, the results showed that the progression of proficiency estimates was relatively stable and interpretable—for example, displaying either a steady linear increase or consistently low values—rather than highly fluctuating. Further work should investigate this more extensively.

Fourth, in Sec. III B, we calculated the overall average RMSE across various σ^i values and based our comparison on the σ^i that yielded the lowest overall average RMSE. This approach reflects practical considerations for implementing the test in classroom settings, where instructors determine σ^i before administration without knowing the amount of proficiency progression in advance. A possible real-world usage scenario would be one in which the teacher chooses either Chain-CAT (with CI) or the pre-post method. In the former (latter) case, the program would automatically select $\sigma^i = 0.3$ ($\sigma^i = 0.6$) to minimize RMSE. We found that, even when sigma is optimized, the minimal RMSE when collateral information is not used is larger than the RMSE when collateral information is used. As such, we do not recommend using Chain-CAT without collateral information. This analysis using session and overall averages may not provide a rigorous comparison, but it can offer a useful reference from a practical perspective that considers the actual conditions of test administration in the classroom. Since our method of analysis is exploratory, there remains room for improvement in such comparative approaches.

Fifth, when expanding the item bank by developing new items, it is important to recognize that evaluating what constitutes a "good" question requires multiple perspectives. The results presented in Fig. 4 illustrate one method of item evaluation, grounded in item response theory. While some items appear to be more discriminating and therefore more informative from the IRT perspective, this approach captures only a portion of what defines item quality. Other important factors—beyond those identified by IRT—may be valued by experts. Thus, to meaningfully assess the quality of new items, additional investigation is needed to determine whether items that perform well under IRT are also regarded as high quality by experts.

2. Improving efficiency further

In addition to the methods we used in this study, there are many options to explore that may further improve the efficiency of the Chain-CAT. First, while we used the 2PL model, the 3PL model is also applicable. In going from 2PL to 3PL, the overall average RMSE values reported in Table II were found to improve by approximately 0.01. It is up to the analyst to decide which model to adopt, considering both the degree of improvement and the increased survey burden associated with the larger sample size required for the 3PL model. Second, although we used the length criterion as the stopping step, an alternative option is to use the precision criterion. This criterion administers as many items as necessary until a prespecified precision is obtained [23]. However, because the test length varies from respondent to respondent when the precision criterion is used, it would be inconvenient to administer the test in a classroom setting, and so at-home administration may be preferable.

The algorithm for the use of collateral information could be improved. In the test step of the Chain-CAT algorithm, we directly used a given student's proficiency estimate; that is, we chose the prior distribution as a normal distribution with $\mu^i = \hat{\theta}_1^i$ for the second test. Another possibility is to predict the initial proficiency estimate of the nth test by a regression model in a similar way to Ref. [34] or another session-series model using the given student's proficiency estimates before the nth test. Another option is to implement the algorithm by utilizing the response time in addition to the response choice of the students [67]. This information might further improve the test efficiency.

3. Possible practical uses of Chain-CAT FCI

One practical application of the Chain-CAT is its integration into ordinary short tests administered after university mechanics lectures. For example, each mini test could consist of eight questions: three problems related to the lecture content and five FCI questions presented in CAT format. Students would be graded only on the first three questions, while the latter five would not contribute to their grade. However, because all questions are presented consecutively on the LMS, it is expected that most students will complete the entire test to ensure their evaluation. This design is not only expected to prevent declines in response rates but also allows instructors to achieve two complementary goals: The first three problems provide a measure of how well students have understood the lecture's specific content. The five FCI questions assess students' conceptual understanding of Newtonian mechanics in general at the time of the lecture. Throughout the semester, instructors can analyze the changes in conceptual understanding for each student.

As an example of how instructors might interpret the Chain-CAT results, we could analyze the impact of specific instructional events, as suggested by Heckler and Sayre [17]. For instance, if providing feedback on homework is followed by a sharp increase in students' conceptual

understanding, this could serve as evidence of the effectiveness of the instructional method. Furthermore, one may find unexpected changes in the proficiency estimate, for example, a decay, as observed in Ref. [17]. The state of understanding that appeared to have temporarily improved was actually unstable and can be interpreted as insufficient understanding. Researchers can use the Chain-CAT as an alternative tool to investigate the nature of student understanding.

The Chain-CAT FCI would be particularly suitable for use in first-year university courses with high school physics as a prerequisite. In Japan, most first-year students who study mechanics at the university level have already encountered mechanics in high school, but their high school education tends to emphasize solving problems using formulas. University courses, especially those employing active learning methods such as tutorials, provide a more appropriate context for analyzing how students' conceptual understanding develops.

4. Development of Chain-CAT feedback system

The following issues need to be addressed to develop the feedback system based on Chain-CAT FCI. First, it is necessary to develop a method for analyzing the progression of individual proficiency. Our simulation results showed a clear progression trend, as shown in Fig. 2, because the statistics were calculated using 10 000 simulated responses at each session. However, for an individual respondent, the progression of proficiency fluctuates much more, which makes it difficult to precisely understand the student's state. (As mentioned above, although the sample size is small, no significant fluctuations were observed in the surveys conducted with actual students. However, it remains possible that such fluctuations could appear in

other contexts.) Therefore, we may need to estimate a student's state based on the analysis of multiple test results, for example, by calculating an average across multiple sessions. It might be appropriate to group the proficiency estimates into several levels, as in Ref. [36].

There are many ways to provide feedback from Chain-CAT to respondents [68]. A simple proficiency estimate alone may not be useful for students; It is more useful to group the proficiency estimates into levels—for example, by simply dividing into four levels based on set thresholds. In Chain-CAT, students can receive feedback in the form of graphs showing their progress, as well as through sentences with encouragement and/or guidance provided by generative artificial intelligence [69]. In considering what feedback to provide, student responses not only on the current test but also on prior tests should be considered. For example, a steadily increasing score across tests (indicating that all is well) might be met with automated feedback such as "I notice your scores are steadily improving. What strategies or habits do you think contributed most to this growth?" On the other hand, for a student with a stagnant score across several tests (indicating that additional effort is required) feedback might take the form of "I see your scores haven't changed much yet. What parts of the test felt the hardest for you this time? Which ones felt more manageable?" In either case, this feedback can promote reflection, metacognition, and self-regulation among students [70]. Developing and implementing these feedback mechanisms will be the focus of our future work.

APPENDIX: ITEM PARAMETER ESTIMATES OF THE FCI USING THE 2PL MODEL

TABLE III. The results of the estimation for the item parameters of the FCI based on the 2PL model ($N = 2712$). The estimates and the standard errors (SEs) for the item parameters are shown.

	$\hat{a}$	SE($\hat{a}$)	$\hat{b}$	SE($\hat{b}$)		$\hat{a}$	SE($\hat{a}$)	$\hat{b}$	SE($\hat{b}$)
Item 1	1.113	0.064	−0.961	0.059	Item 16	1.231	0.066	−0.710	0.048
Item 2	0.844	0.052	−0.425	0.056	Item 17	1.453	0.071	0.525	0.041
Item 3	0.923	0.068	−1.915	0.123	Item 18	1.568	0.074	−0.069	0.036
Item 4	1.064	0.058	0.586	0.051	Item 19	1.059	0.059	−0.533	0.049
Item 5	1.453	0.070	0.225	0.038	Item 20	1.233	0.063	−0.343	0.042
Item 6	0.900	0.060	−1.365	0.088	Item 21	0.980	0.055	0.471	0.052
Item 7	0.942	0.058	−0.906	0.064	Item 22	1.145	0.060	−0.125	0.043
Item 8	1.201	0.069	−1.102	0.060	Item 23	1.351	0.066	0.276	0.040
Item 9	1.140	0.061	−0.444	0.046	Item 24	1.254	0.067	−0.769	0.049
Item 10	1.571	0.096	−1.458	0.065	Item 25	1.893	0.086	0.156	0.033
Item 11	1.327	0.065	0.229	0.040	Item 26	1.966	0.093	0.567	0.035
Item 12	1.156	0.079	−1.777	0.098	Item 27	0.881	0.054	−0.592	0.058
Item 13	2.916	0.139	0.061	0.028	Item 28	1.188	0.064	−0.634	0.047
Item 14	1.135	0.059	0.096	0.043	Item 29	0.182	0.049	−6.645	1.799
Item 15	0.591	0.047	0.716	0.087	Item 30	2.011	0.093	0.435	0.034

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30,** 141 (1992).

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

[3] A. Pawl, Using force concept inventory data to develop impactful class activities, Phys. Teach. **58,** 94 (2020).

[4] B. R. Wilcox, S. J. Pollock, and D. R. Bolton, Retention of conceptual learning after an interactive introductory mechanics course, Phys. Rev. Phys. Educ. Res. **16,** 010140 (2020).

[5] E. Burkholder, C. Walsh, and N. Holmes, Examination of quantitative methods for analyzing data from concept inventories, Phys. Rev. Phys. Educ. Res. **16,** 010141 (2020).

[6] Y. Xiao, G. Xu, J. Han, H. Xiao, J. Xiong, and L. Bao, Assessing the longitudinal measurement invariance of the Force Concept Inventory and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **16,** 020103 (2020).

[7] A. Reinhard, A. Felleson, P. C. Turner, and M. Green, Assessing the impact of metacognitive postreflection exercises on problem-solving skillfulness, Phys. Rev. Phys. Educ. Res. **18,** 010109 (2022).

[8] C. Wheatley, J. Wells, D. E. Pritchard, and J. Stewart, Comparing conceptual understanding across institutions with module analysis, Phys. Rev. Phys. Educ. Res. **18,** 020132 (2022).

[9] A. Morley, J. M. Nissen, and B. V. Dusen, Measurement invariance across race and gender for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **19,** 020102 (2023).

[10] E. Christman, P. Miller, and J. Stewart, Beyond normalized gain: Improved comparison of physics educational outcomes, Phys. Rev. Phys. Educ. Res. **20,** 010123 (2024).

[11] M. Jariwala, J.-S. S. White, B. V. Dusen, and E. W. Close, In-class vs online administration of concept inventories and attitudinal assessments, presented at PER Conf. 2016, Sacramento, CA, 10.1119/perc.2016.pr.039.

[12] J. M. Nissen, M. Jariwala, E. W. Close, and B. V. Dusen, Participation and performance on paper- and computer-based low-stakes assessments, Int. J. STEM Educ. **5,** 21 (2018).

[13] B. V. Dusen, M. Shultz, J. M. Nissen, B. R. Wilcox, N. G. Holmes, M. Jariwala, E. W. Close, H. J. Lewandowski, and S. Pollock, Online administration of research-based assessments, Am. J. Phys. **89,** 7 (2021).

[14] R. S. Siegler and K. Crowley, The microgenetic method: A direct means for studying cognitive development, Am. Psychol. **46,** 606 (1991).

[15] D. Kuhn, Microgenetic study of change: What has it told us?, Psychol. Sci. **6,** 133 (1995).

[16] E. C. Sayre and A. F. Heckler, Peaks and decays of student knowledge in an introductory EM course, Phys. Rev. ST Phys. Educ. Res. **5,** 013101 (2009).

[17] A. F. Heckler and E. C. Sayre, What happens between pre- and post-tests: Multiple measurements of student understanding during an introductory physics course, Am. J. Phys. **78,** 768 (2010).

[18] E. C. Sayre, S. V. Franklin, S. Dymek, J. Clark, and Y. Sun, Learning, retention, and forgetting of Newton's third law throughout university physics, Phys. Rev. ST Phys. Educ. Res. **8,** 010116 (2012).

[19] R. Brock and K. S. Taber, The application of the microgenetic method to studies of learning in science education: Characteristics of published studies, methodological issues and recommendations for future research, Stud. Sci. Educ. **53,** 45 (2017).

[20] R. Brock and K. S. Taber, Making claims about learning: A microgenetic multiple case study of temporal patterns of conceptual change in learners' activation of force conceptions, Int. J. Sci. Educ. **42,** 1388 (2020).

[21] R. Brock, K. S. Taber, and D. M. Watts, Assembly required: A microgenetic multiple case study of four students' assemblages when learning about force, Int. J. Sci. Educ. **46,** 1027 (2024).

[22] Elements of Adaptive Testing, edited by W. J. van der Linden and C. A. Glas (Springer, New York, 2010).

[23] D. Magis, D. Yan, and A. A. von Davier, Computerized Adaptive and Multistage Testing with R (Springer, New York, 2017).

[24] International vocabulary of metrology—Basic and general concepts and associated terms (VIM) 3rd edition, Technical Report No. JCGM 200:2012, Joint Committee for Guides in Metrology, 2012.

[25] C. N. Mills and M. Steffen, The GRE computer adaptive test: Operational issues, in Computerized Adaptive Testing: Theory and Practice (Springer, Netherlands, 2000), pp. 75–99.

[26] College Board, Assessment Framework for the Digital SAT Suite, version 3.01 (College Board, New York, 2024).

[27] A. Frey, C. König, and A. Fink, A highly adaptive testing design for PISA, J. Educ. Measure. **62,** 415 (2023).

[28] 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, Phys. Rev. Phys. Educ. Res. **14,** 020110 (2018).

[29] M. A. Samsudin, T. S. Chut, M. E. Ismail, and N. J. Ahmad, A calibrated item bank for computerized adaptive testing in measuring science TIMSS performance, Eurasia J. Math. Sci. Technol. Educ. **16,** em1863 (2020).

[30] M. D. Linderman, S. A. Suckiel, N. Thompson, D. J. Weiss, J. S. Roberts, and R. C. Green, Development and validation of a comprehensive genomics knowledge scale, Public Health Genomics **24,** 291 (2021).

[31] R. Zakwandi, E. Istiyono, and W. S. B. Dwandaru, A two-tier computerized adaptive test to measure student computational thinking skills, Educ. Inf. Technol. **29,** 8579 (2024).

[32] J. Morphew, A. Mehrabi, B. V. Dusen, and J. Nissen, AI-based concept inventories: Using cognitive diagnostic computer adaptive testing in lasso for classroom assessment,

in *Paper Presented at 2024 ASEE Annual Conference and Exposition* (2024), 10.18260/1-2–46534.

[33] V. Le, B. V. Dusen, J. M. Nissen, X. Tang, Y. Zhang, H. H. Chang, and J. W. Morphew, Mechanics cognitive diagnostic: Mathematics skills tested in introductory physics courses, *presented at PER Conf. 2024, Boston, MA*, 10.1119/perc.2024.pr.Le.

[34] W. J. van der Linden, Empirical initialization of the trait estimator in adaptive testing, Appl. Psychol. Meas. **23**, 21 (1999).

[35] B. P. Veldkamp, M. Matteucci, and T. J. Eggen, Computerized adaptive testing in computer assisted learning?, Commun. Comput. Inf. Sci. **126**, 28 (2011).

[36] Y. Choi and C. McClenen, Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks, Appl. Sci. **10**, 8196 (2020).

[37] A. C. Yang, B. Flanagan, and H. Ogata, Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning, Comput. Educ. **3**, 100104 (2022).

[38] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2024), https://www.R-project.org/.

[39] R. P. Chalmers, MIRT: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).

[40] D. Magis and G. Raiche, Random generation of response patterns under computerized adaptive testing with the R package catR, J. Stat. Softw. **48**, 1 (2012).

[41] D. Magis and J. R. Barrada, Computerized adaptive testing with R: Recent updates of the package catR, J. Stat. Softw. **76**, 1 (2017).

[42] C. DeMars, *Item Response Theory* (Oxford University Press, New York, 2012).

[43] J. Yasuda, N. Mae, M. M. Hull, and M. Aki Taniguchi, Optimizing the length of computerized adaptive testing for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **17**, 010115 (2021).

[44] J. Yasuda, M. M. Hull, and N. Mae, Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **18**, 010112 (2022).

[45] W. M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, Appl. Psychol. Meas. **8**, 125 (1984).

[46] A. Maydeu-Olivares, Goodness-of-fit assessment of item response theory models, Meas. Interdiscip. Res. Perspect. **11**, 71 (2013).

[47] G. Schwarz, Estimating the dimension of a model, Ann. Stat. **6**, 461 (1978).

[48] J. Yasuda and M. M. Hull, Balancing content of computerized adaptive testing for the force concept inventory, *presented at PER Conf. 2021, Virtual Conference*, 10.1119/perc.2021.pr.Yasuda.

[49] G. G. Kingsbury and A. R. Zara, Procedures for selecting items for computerized adaptive tests, Appl. Meas. Educ. **2**, 359 (1989).

[50] N. A. Thompson and D. J. Weiss, A framework for the development of computerized adaptive tests, Pract. Assess. Res. Eval. **16**, 1 (2011).

[51] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, 4th ed. (Wiley, New York, 2012).

[52] M. H. Dancy and R. Beichner, Impact of animation on assessment of conceptual understanding in physics, Phys. Rev. ST Phys. Educ. Res. **2**, 010104 (2006).

[53] S. O. Popp and J. Jane, Can assessment of student conceptions of force be enhanced through linguistic simplification? presented at the American Educational Research Association 2009, San Diego, CA, http://modeling.asu.edu/R&E/OsbornPopp-SFCI-aera09.pdf.

[54] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 020109 (2010).

[55] L. McCullough and D. E. Meltzer, Differences in male/female response patterns on alternative-format versions of the Force Concept Inventory, presented at PER Conf. Proc. 2001, Rochester, NY, 10.1119/perc.2001.pr.013.

[56] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[57] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).

[58] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the force and motion conceptual evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[59] P. Eaton, B. Frank, and S. Willoughby, Detecting the influence of item chaining on student responses to the Force Concept Inventory and the force and motion conceptual evaluation, Phys. Rev. Phys. Educ. Res. **16**, 020122 (2020).

[60] D. Huffman and P. Heller, What does the force concept inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[61] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[62] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[63] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[64] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance on the Force Concept Inventory using factor analysis, Phys. Rev. Phys. Educ. Res. **13**, 010103 (2017).

[65] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).

[66] R. P. Chalmers, Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications, J. Stat. Softw. **71**, 1 (2016).

[67] W. J. van der Linden, Using response times for item selection in adaptive testing, J. Educ. Behav. Stat. **33**, 5 (2008).

[68] K. Duss, *Formative Assessment and Feedback Tool* (Springer Fachmedien, Wiesbaden, 2020).

[69] S. Küchemann, S. Steinert, J. Kuhn, K. Avila, and S. Ruzika, Large language models—valuable tools that require a sensitive integration into teaching and learning physics, Phys. Teach. **62**, 400 (2024).

[70] I. Clark, Formative assessment: Assessment is for self-regulated learning, Educ. Psychol. Rev. **24**, 205 (2012).