Designing a computerized adaptive testing chain for the Force Concept Inventory

Jun-ichiro Yasuda (he/him/his)
Center for the Studies of Higher Education, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Michael M. Hull (he/him/his)
Department of Physics, University of Alaska Fairbanks, 1930 Yukon Dr, Fairbanks, Alaska 99775, USA

Kentaro Kojima (he/him/his)
Faculty of Arts and Science, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan

The use of computer adaptive testing (CAT)-based assessment tests has inherent issues associated with the pre- and post-paradigm, such as the limited ability to observe the progression of student conceptual understanding throughout the course. To address these issues, we propose increasing the frequency of CAT-based assessments during the course, while reducing the test length per administration, thus decreasing the total number of test items during the course. The feasibility of this idea depends on how far the test length per administration can be reduced. To reach this goal, we designed a CAT algorithm, which we call Chain-CAT. This algorithm sequentially links the results of each CAT administration using collateral information. We analyzed the advantages of this algorithm by numerical simulations. Although preliminary, we found that collateral information significantly improved the test efficiency, and the total test length could be shorter than the pre-post method.

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

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 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 the reference values in the literature) to determine 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 the test. 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.

Despite the benefits, some teachers are reluctant to allocate their class time for the assessment. One reason is that each pre- and post-administration of the FCI can take approximately 40 minutes (including the time needed to orient students to the survey), which can reduce student learning time and overwhelm the teachers' schedules that is already crowded with curriculum requirements.

To shorten the test time, Yasuda *et al.* [3, 4] suggested the use of computerized adaptive testing (CAT). 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 next answers an easier item. In this way, high (low) proficiency students do not need to answer items that are too easy (difficult) for them; thereby, the test length can be significantly shortened in comparison to standard test administration [5, 6]. According to the simulation results of Yasuda *et al.* [3, 4], the CAT-based FCI under typical conditions can reduce the test length to about half of the full-length FCI without compromising accuracy and precision (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 [7].) Because of its efficiency, CAT is becoming widely used, for example, with the Graduate Record Exam (GRE) [8], with PISA [9], and recently in science education research [10–12].

Although CAT provides possible solutions to reduce test time, there are still specific issues associated with the pre-post paradigm. First, the pre-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 their conceptual understanding throughout the duration of the course. Second, students may see little benefit in their 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. This situation can make students feel burdened and may decrease their engagement and motivation to take the assessment.

To address these issues associated with the pre-post paradigm, we propose increasing the frequency of CAT-based assessments during the course, while reducing the test length per administration, thus decreasing the total number of test items during the course. This idea was inspired by the microgenetic method [13–19], which includes frequent quantitative and/or qualitative measurements. This method involves sampling data at a frequency that is assumed to be high compared to the rate of change of the phenomenon of interest and collecting data for the entirety of the period of change [18], allowing teachers to monitor students' understanding progression. Although these methods have been limited in practice to a small number of students, Sayre and Heckler [15, 16] extended the method to the multiple-choice tests administered to large numbers of students, by posing simple conceptual questions to separate randomly selected groups of students.

Our idea of frequent administration of short CAT-based assessments can be used as a form of formative assessment. Providing automated feedback according to each student's set of responses is expected to increase the usefulness of the survey for students and reduce their sense of burden.

The feasibility of the above idea depends on how far the test length per administration can be reduced without compromising accuracy and precision. For CAT where the item bank consists of the 30 FCI items, 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 10 administrations of the CAT-based FCI, it is then desired that the test length per administration would be less than 6 items. To reach this goal, we utilize a CAT algorithm that takes advantage of collateral information [5, 20]. Collateral information is the relevant empirical information on the respondents, for example, age, grade, or previous test scores. This information can be used to select the first item in CAT and to specify the prior distribution for the proficiency estimation based on the Bayesian method [20]. In so doing, we can accelerate the convergence of the proficiency estimation during the test administration, hence improving test efficiency [5]. Specifically, as we describe below, we use the proficiency estimate of each respondent in a test for selecting items and estimating respondent proficiency level in the next test for the respondent. Since this CAT algorithm sequentially links the test result of each administration, we call this algorithm *Chain-CAT*.

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 computerized adaptive, and each test length is relatively small (e.g., from 5 to 10 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

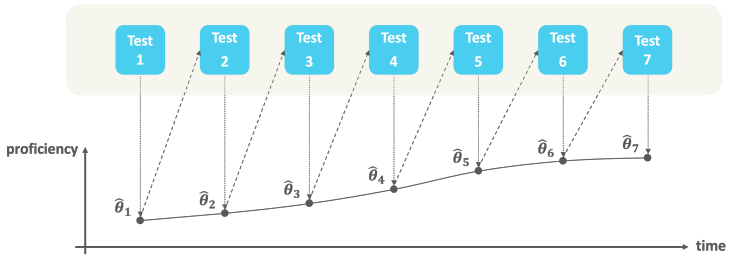

FIG. 1. Illustration of Chain-CAT algorithm for a student taking the tests for a course with seven administrations of Chain-CAT.

sequentially from the next test to the final test (Test 7).

The objective of this study is to analyze the advantages of using collateral information in the Chain-CAT algorithm. Specifically, our research question is: How much does collateral information improve the test efficiency when it is used to sequentially link the results of each CAT administration? This work is the first step to examine the feasibility of the Chain-CAT version of the FCI (Chain-CAT FCI).

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 analyzing its efficiency. In Sec. III, we describe the preliminary simulation results. Finally, in Sec. IV, we summarize this study and show the prospects.

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 [21]. In this model, the probability of a correct response from the ith respondent on item j is given by

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

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 used the estimated item parameters of the FCI determined by Yasuda *et al.* [3, 4]. In the process, they administered the full-length paper-and-pencil (in-class) FCI to 2882 university students from April 2015 to April 2018. The assumptions of unidimensionality, overall local independence, and goodness of fit were confirmed, which validates the use of the 2PL model in our analysis.

B. Basic computerized adaptive testing process

The CAT process consists of four successive steps [6]: (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 used criterion to select the first item is the maximum Fisher information (MFI) criterion [6]. 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. When nothing is known about the respondent, the Fisher information of the item is calculated using the mean proficiency value of the prior population (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, we used the expected a posteriori (EAP) method as in [3, 4]. 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 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 for 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. There are three stages where we can utilize collateral information: the initial step, the test step, and the final step of the testing process. In what follows, we explain how we implemented collateral information for each stage.

a. Initial step In this step, the first item is selected and administered to a respondent, as we described above. 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 be zero to have the scale be centered on respondents [6], as we described above.

At the beginning of the second test, in the Chain-CAT algorithm, the Fisher information can be calculated utilizing the proficiency estimate of the first test for a given student as collateral information to improve the test efficiency for that student. Similarly, 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.

Generally, for each test, the farther the initial proficiency estimate is from the true proficiency of the respondent, the slower the algorithm converges [20]. To decrease this gap, we directly used a given student's proficiency estimate from the $(n-1)$th test as the initial proficiency estimate of the nth test (as in [22]).

b. 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 in turn is proportional to the product of the likelihood function and a prior distribution of the proficiency $g(\theta^i)$ for an ith respondent [6]. We used a model with a normal distribution, thereby 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.

On the first test, when we know nothing about the respondents beforehand, a common choice of the prior distribution is the standard normal distribution, with $\mu^i = 0$ and $\sigma^i = 1$, namely, $g(\theta^i) \sim \mathcal{N}(0, 1)$ [6]. On the second test, we can utilize the first test proficiency estimate of an ith respondent, $\hat{\theta}_1^i$ as collateral information for the prior distribution. Specifically, in a similar way above, we directly used a given student's proficiency estimate; namely, we chose the prior distribution as the normal distribution with $\mu^i = \hat{\theta}_1^i$. Similarly, the $(n-1)$th test proficiency estimate of an ith respondent, $\hat{\theta}_{n-1}^i$ was utilized as collateral information and the normal distribution with $\mu^i = \hat{\theta}_{n-1}^i$ is chosen as the prior distribution for the nth test.

In the model directly using the proficiency estimate, the standard deviation of the prior distribution cannot be estimated. Generally, unless reliable collateral information about the examinee is available, the prior distribution should be chosen to be low informative (namely, with a relatively large standard deviation) [5].

c. Final step We used collateral information also for the final proficiency estimation using the EAP method as just described for the test step.

D. Numerical simulation procedure

We conducted a numerical simulation to analyze the efficiency of the Chain-CAT FCI and to search for the optimal algorithm. The outline of the procedure is as follows.

First, to analyze the efficiency of the Chain-CAT FCI as generally as possible, we assumed various progression patterns of the true-value proficiency θ. Specifically, we assumed the following progression models for $\theta(t)$, where t represents time: a) a stationary model in which θ is constant with respect to time, b) a linear model in which θ increases linearly with respect to time, and c) a step model in which θ increases significantly only at a certain point in time. 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). We supposed that most of the student's progressions are one of these patterns or a combination of them. Second, we generated response data based upon a Monte Carlo simulation, which is commonly used in the development of CAT [23]. Specifically, we generated the responses for Chain-CAT using the catR package. In this analysis, we generated 1000 response data for each time point, where the dispersion of the estimated proficiency mean is almost negligible. Third, we represented the accuracy and precision by the bias and standard error, respectively. Then, we summarized these measures in terms of the root-mean-square error (RMSE) [24].

We calculated the accuracy and precision of the Chain-CAT FCI and compared them to those obtained when the FCI is conducted in the ordinary pre-post method. The latter was calculated by having all 30 FCI items administered both in the pretest and post-test and estimating student proficiency using item response theory in the same model as above (2PL model). Specifically, we calculated the RMSE of the pre-post method and used this as a reference value with which to examine the efficiency of the Chain-CAT FCI.

III. RESULT

Figure 2 shows the simulation results assuming the linear model for the true proficiency progression. Collateral information is not implemented in Fig. 2 (a) and is implemented in Fig. 2 (b). The figures include the graphs for comparing the test length of each administration $L = 5, 10, 15$ items. The graphs include the simulation results for nine time-points from t_1 to t_9, where the true proficiency level increases from $\theta(t_1) = 0$ to $\theta(t_9) = 0.8$ (see the purple lines in the graphs). The standard deviation σ^i of the prior distribution in the Chain-CAT algorithm was fixed at 0.5. For each time point, we generated 1000 responses 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 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 (the green lines). The y-axis in Fig. 2 reports these statistics. In Fig. 2 (b), the standard error is much larger than the bias, so the standard error values and the RMSE values are

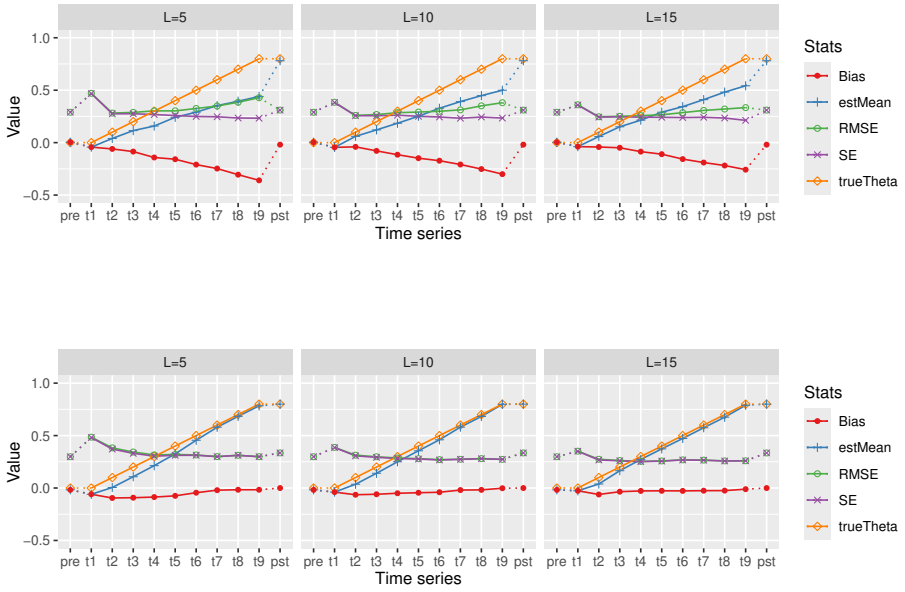

FIG. 2. Simulation results for a linear model comparing the test length of each administration $L = 5, 10, 15$ items. (a) Top: without collateral information (Simple CAT). (b) Bottom: with collateral information (Chain-CAT), which is used from t_2 to t_9. "Pre" and "pst" show the results of the 30-item pretest and post-test, which were used as reference values for comparing the Chain CAT-FCI.

very close, and the SE and the RMSE graphs almost overlap. In the graphs, the reference values are shown at $t =$ pre and $t =$ pst on the horizontal axis, where the statistics were calculated using the true proficiency levels at $t = t_1$ and $t = t_9$ and administering the whole FCI items (30 items) in both cases.

From the above figures, we can observe the following two things. First, in this case, the collateral information significantly improved the accuracy and precision. Without the collateral information, the deviation between the true theta and the estimated theta became large as the number of tests increased. Specifically, on average from t_1 to t_9, the bias was reduced by (64.3%, 73.4%, 77.3%) for $L = (5, 10, 15)$, respectively. Second, by visual inspection, the accuracy and precision of Chain-CAT seem to become comparable to the reference value of $t =$ pst after conducting a certain number of tests. Specifically, the RMSE when $L = 5$ became comparable to the reference value at $t = t_3$, the RMSE when $L = 10$ and $L = 15$ became comparable at $t = t_2$. These results suggest that the total test length can be shortened by increasing the test length for the first part of the test and then shortening the test length after sufficient accuracy and precision are achieved. For example, if we set $L = 10$ at $t = t_1$ and $t = t_2$, then $L = 5$ after $t = t_2$, the accuracy and precision of $t_2 \leq t \leq t_9$ would be comparable to the reference value, and the total test length is 55 items, which is shorter than the pre-post method (60 items). Although preliminary, we found similar results for the other proficiency progression models and conditions.

IV. CONCLUSIONS

To address the issues associated with the pre-post paradigm, we proposed increasing the frequency of CAT-based assessments during the course, while reducing the test length per administration, thus decreasing the total number of test items during the course. This idea was realized as the Chain-CAT algorithm, which sequentially links the results of each CAT administration using collateral information. By conducting numerical simulations, although preliminary, we found that the Chain-CAT algorithm significantly improved the test efficiency, and the total test length could be shorter than the pre-post method. Our results also implied that the total test length can be shortened by increasing the test length for the first part of the test and then shortening the test length after sufficient accuracy and precision are achieved.

To use the Chain-CAT FCI as a practical test, it is necessary to consider the constraints for the item selection (content balancing and item exposure), which improves the validity of the Chain-CAT FCI but could reduce the test efficiency. It is also necessary to construct an item bank with desirable properties (unidimensionality etc.). Our future work will attend to these analyses.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP 22H01061.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **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, American Journal of Physics **66**, 64 (1998), doi: 10.1119/1.18809.

[3] J. ichiro Yasuda, N. Mae, M. M. Hull, and M. aki Taniguchi, Optimizing the length of computerized adaptive testing for the force concept inventory, Physical Review Physics Education Research **17**, 010115 (2021).

[4] J. ichiro Yasuda, M. M. Hull, and N. Mae, Improving test security and efficiency of computerized adaptive testing for the force concept inventory, Physical Review Physics Education Research **18**, 010112 (2022).

[5] W. J. van der Linden and C. A. W. Glas, *Elements of adaptive testing* (Springer, 2010).

[6] D. Magis, D. Yan, and A. A. von Davier, *Computerized adaptive and multistage testing with R* (Springer, 2017).

[7] International vocabulary of metrology - basic and general concepts and associated terms (vim) 3rd edition (2012).

[8] C. N. Mills and M. Steffen, The gre computer adaptive test: Operational issues bt - computerized adaptive testing: Theory and practice (Springer, 2000).

[9] K. Yamamoto, H. J. Shin, and L. Khorramdel, Introduction of multistage adaptive testing design in pisa 2018, OECD Education Working Papers (2019).

[10] J. W. Morphew, J. P. Mestre, H.-A. Kang, H.-H. Chang, and G. Fabry, Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course, Physical Review Physics Education Research **14**, 020110 (2018).

[11] 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 Journal of Mathematics, Science and Technology Education **16**, em1863 (2020).

[12] 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 , 1 (2021).

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

[14] D. Kuhn, Microgenetic study of change: What has it told us?, https://doi.org/10.1111/j.1467-9280.1995.tb00322.x **6**, 133 (1995).

[15] E. C. Sayre and A. F. Heckler, Peaks and decays of student knowledge in an introductory em course, Physical Review Special Topics - Physics Education Research **5**, 013101 (2009).

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

[17] 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, Studies in Science Education **53**, 45 (2017).

[18] 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, International Journal of Science Education **42**, 1388 (2020).

[19] 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, Physical Review Special Topics - Physics Education Research **8**, 010116 (2012).

[20] W. J. V. D. Linden, Empirical initialization of the trait estimator in adaptive testing, Applied Psychological Measurement **23**, 21 (1999).

[21] C. DeMars, *Item response theory* (Oxford University Press, 2012).

[22] Q. Xie, The impact of collateral information on ability estimation in an adaptive test battery, University of Iowa 10.17077/etd.njvy-42a6 (2019).

[23] N. A. Thompson and D. J. Weiss, A framework for the development of computerized adaptive tests, Practical Assessment, Research and Evaluation **16**, 1 (2011).

[24] J. S. Bendat and A. G. Piersol, *Random data: analysis and measurement procedures*, 4th ed. (Wiley, 2012).