Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory

Jun-ichiro Yasuda[*]
Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan

Michael M. Hull
Austrian Educational Competence Centre, Division of Physics, University of Vienna, Vienna 1090, Austria

Naohiro Mae
Research Center for Nuclear Physics, Osaka University, Ibaraki, Osaka 567-0047, Japan

(Received 12 October 2021; accepted 27 January 2022; published 17 February 2022)

This paper presents improvements made to a computerized adaptive testing (CAT)-based version of the FCI (FCI-CAT) in regards to test security and test efficiency. First, we will discuss measures to enhance test security by controlling for item overexposure, decreasing the risk that respondents may (i) memorize the content of a pretest for use on the post-test or (ii) share information about the items with their classmates who take the assessment later. Second, we will discuss measures to enhance test efficiency, so that a shorter test length can yield a desired accuracy and precision of the measurement. Specifically, we utilized collateral information in the form of a pretest proficiency estimate of each respondent for selecting items and estimating respondent proficiency level in the post-test. To shorten the total testing time further, we also allowed the test lengths to be different for the pre- and post-test. To analyze how these improvements affect the accuracy and precision (which we measure in terms of root-mean-square error) of Cohen's d, we conducted a Monte Carlo simulation and a *post hoc* simulation. Then, we calculated the minimal test length of the FCI-CAT whose accuracy and precision are equivalent to that of the paper-and-pencil version of the FCI. Consequently, we obtained the following three findings: (i) By using collateral information, we can achieve the accuracy and precision of the full-length FCI with fewer items via the FCI-CAT. (ii) For a class size of 40, we can control for test security while still reducing the sum of the pre- and post-test lengths of the FCI-CAT to a total of 33 items (17 items on the pretest and 16 items on the post-test), thereby reducing the testing time to 55%. (iii) If one's goal is to maximize test efficiency, the pretest length should be slightly larger than the post-test length. On the other hand, if the goal is to maximize test security, the pretest length should be smaller and the post-test length should be larger. If one desires a balance of these two goals, it would be reasonable to choose equal pre- and post-test lengths.

DOI: 10.1103/PhysRevPhysEducRes.18.010112

I. INTRODUCTION

Research-based assessments play an important role in informing physics teachers and education researchers about what students learn in physics courses. One of the most widely used research-based assessments is the Force Concept Inventory (FCI) [1]. The FCI probes student conceptual understanding of Newtonian mechanics, particularly regarding the concept of force. The test has 30 items with five choices, and students typically take 20–30 min to complete the test. The FCI has undergone a rigorous validation process [2–11] and is used internationally with a wide variety of students [12–15]. As such, it allows for a standardized comparison of student understanding on the concept of force. By administering the FCI both before and after instruction, we can measure the effects of that instruction in terms of improvement of students' scores [16–23].

Many instructors feel pressure to cover a demanding expanse of content by the end of the semester, and they are likely to be reluctant to carve time out of their crowded schedules to administer the assessment [24]. To reduce the test time, Han *et al.* [25] divided the FCI into two half-length tests which contain different subsets of the original FCI. To avoid using class time for assessments, some instructors administer the assessment via online platforms which enables students to complete the assessment outside

[*] phys.cat.collaboration@gmail.com

of class [24,26–30]. Although this preserves in-class time, it does not solve the problem of consuming student time, time that students could otherwise spend doing additional homework or independent study. Moreover, administering an ungraded survey online outside of class can decrease response rate and compromise test security [31–33].

Recently, we [34] suggested the use of computerized adaptive testing (CAT) to reduce the test time. CAT is the practice of using a computer to administer successive items in the test to match the current estimate of the student's proficiency. 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 in comparison to standard test administration [35,36]. Because of its efficiency, CAT is becoming widely used, for example, with the Graduate Record Exam (GRE) [37], with PISA [38], and, recently, in science education research [39–41].

When developing a computerized adaptive test version of the FCI (FCI-CAT), one of the key questions is how much can we shorten the test length without excessively compromising the accuracy and precision of the instrument? (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 [42].) This question can be rephrased as "how efficient is the FCI-CAT?", where *test efficiency* is defined by the minimal test length yielding a desired accuracy and precision of the measurement [35] (the shorter the test length, the more efficient it is).

Previously, we [34] focused on the accuracy and precision of the standardized mean difference [43], a statistic to quantify the pre- and post-group difference. Our work focuses on the group difference in keeping with the previous study by Han *et al.* [25]. Whereas they used the average normalized gain as a metric, we [34] use the standardized mean difference to check for consistency with the analysis of Demars [44], who analyzed accuracy and precision in the context of item response theory. (For details, see the introduction of Ref. [34].) Based on simulation studies, we found that the test length of the FCI-CAT may be reduced to 15–19 items with an accompanying decrease in accuracy and precision of 5%–10% from what is obtained with the full length FCI.

The FCI-CAT can be validated and improved in various means, for example, in terms of *test security* [45], which are commonly considered in developing CAT [36,46]. When students complete an assessment at different points in time, there is a risk that respondents may memorize the content and share it with their classmates who take the assessment later. Furthermore, when students take the same assessment as a post-test, memorization of the material from the pretest may result in inflated improvement that does not reflect actual learning. It was shown that the test-retest memory effect of the FCI is negligible after 3 weeks [47]; however, if researchers want to study shorter-term learning gains to analyze learning progressions [48–50], for example, via the microgenetic approach [51], it is necessary to address this issue carefully. Han *et al.* [25] accounted for the test-retest memorization effect by dividing the FCI into two half-length tests that had only three items in common (for test equating). However, since the tests are linear (question 2 always comes after question 1, etc.), all items are exposed to all respondents, and so there remains the risk of classmates leaking information about the items to their peers. In CAT, item exposure is controlled by the testing algorithm (for example, randomizing the item selection at the early stages of CAT), several candidates for which we consider below.

Test security is an important aspect of instrument validity; however, implementing an algorithm taking into account this issue may come at a cost in terms of test efficiency [52]. When test security is not controlled for, CAT generally begins with the more *informative* items, where "informative" means that a greater precision of the proficiency estimate will be obtained (see technical description below). Although this improves the test efficiency, these items become overexposed, and it increases the chance that these items will be memorized and shared with other respondents prior to taking the assessment. When an algorithm to reduce overexposure is implemented, these informative items are less likely to be selected, and this results in a slowing of the convergence of proficiency estimates: longer test lengths become necessary.

In order to compensate for this decrease in test efficiency, we can utilize *collateral information* [35,53–55]. 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 [56]. In so doing, we can accelerate the convergence of the estimates during the test administration, hence improving test efficiency [35]. Specifically, as we describe below, we use the pretest proficiency estimate of each respondent for selecting items and estimating respondent proficiency level in the post-test (Fig. 1). There is a second benefit to the use of collateral information. If the first item is selected based on collateral information, it results in a variable entry point to the item pool, and hence offers a more even exposure of assessment items [35], which improves test security (details are described below).

To explore the possibility of improving test efficiency and/or security further, we considered the case when the

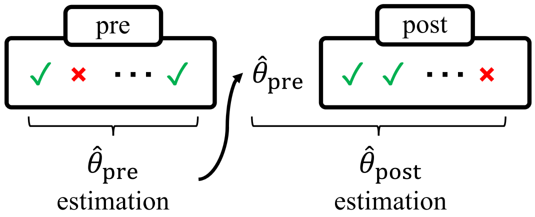

FIG. 1. Illustration of the use of collateral information for proficiency estimation on the post-test.

test length is different for the pre- and post-test. As we will show below, having asymmetric test lengths like this allows us to further reduce the sum of the pre- and post-test lengths without compromising the accuracy and precision.

The objective of this paper is to significantly improve the FCI-CAT by addressing the above-mentioned issues of test security and test efficiency so as to find the optimal test length of the FCI-CAT. Specifically, our research questions are the following: (i) How much do test security algorithms decrease item exposure of the FCI-CAT? (ii) How much do test security algorithms compromise the accuracy and precision for a given FCI-CAT length? (iii) How much does collateral information improve test efficiency of the FCI-CAT? (iv) Using collateral information, what is the minimal test length of the FCI-CAT whose accuracy and precision are equivalent to that of the full length FCI? (v) What is the optimal test length of the FCI-CAT considering both test security and test efficiency?

The remainder of this paper is organized as follows. In Sec. II, we describe the mathematical model we employed in the FCI-CAT, our CAT settings, our approach to analyze the security and efficiency, and the simulation procedures. In Sec. III, we present the results of our analysis. Finally, in Sec. IV, we summarize this study and discuss the limitations of our research and future prospects for it.

All of our analyses were conducted using R [57]. In addition to the basic package of R, the item parameters of the FCI were calibrated using the package MIRT [58] and the simulations of the FCI-CAT were conducted using the package CATR [59,60].

II. METHODOLOGY

A. Item response theory

1. Model

CAT employs item response theory (IRT) as the psychometric model. Models of IRT describe the relationship between the latent trait measured by the instrument and the response to an individual item [61]. Although there are various IRT models to choose from, we used the three-parameter logistic (3PL) model to facilitate comparison with our previous study [34]. In the model, the probability of a correct response of the ith respondent on item j is given by

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

where 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 Eq. (1), 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. The third parameter g_j represents the probability that a respondent would answer an item correctly by guessing.

2. Calibration and model validation

We use the item parameter estimates for a_j, b_j, and g_j calibrated in our previous study [34]. In this calibration process, we administered the full-length paper-and-pencil (in-class) FCI to 2882 Japanese university students from April 2015 to April 2018. The respondents were students at the beginning of introductory physics courses at one public university and four private universities. All five of these schools are middle-rank universities in Japan. From this dataset, we removed aberrant responses to be left with 2712 valid responses. Most of the respondents were first-year students of the department of science or the department of technology from a mix of calculus-based and algebra-based courses. We confirmed that the standard errors of the parameter estimates are not significant (see Table III in Ref. [34]).

In order to validate the model, we confirmed in our previous study [34] that the assumptions of unidimensionality, overall local independence, and goodness of fit are satisfied for the 3PL model. Specifically, we examined the unidimensionality of the FCI via a principal component analysis with the tetra-choric correlation matrix [61], we evaluated the local independence assumption by using Yen's Q_3 statistic [62], and we evaluated the goodness of fit of the 3PL model to the response data with the standardized root mean square residual (SRMSR) [63].

B. Testing process

We model our survey respondents as having a true proficiency level. In CAT, the testing algorithm estimates this proficiency level based upon the respondent's answers to prior items, and this estimate is updated with each item responded to. The next item administered is based upon this estimated proficiency and the calibrated item parameters of the items available. This process can be conceptualized as consisting of four successive steps [36]: (i) initial step, (ii) test step, (iii) stopping step, and (iv) final step. Our settings for the four steps follow.

(i) *Initial step*.—In this 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 [36]. The MFI criterion calls for selecting the most informative item (the item with the largest Fisher information) for the respondent based upon the current estimate of the proficiency. When nothing is known about the respondent (as is often the case when the first item is chosen), the Fisher information of the item is calculated using the mean proficiency value of the prior population. In the pretest, as is commonly done [36], we set the prior population mean proficiency value to be zero to have the scale be centered on respondents. In the post-test, as we describe below, there is an option to use collateral information in the form of the pretest proficiency estimate of a student as the empirical prior to calculate the Fisher information of the item. When considering test security, instead of the MFI criterion, we used other methods to select the first item, as we describe below.

(ii) *Test step*.—In this step, the proficiency of the respondent is estimated using the current set of item responses and the next item is selected to be administered. As in our previous study [34], we used the expected *a posteriori* (EAP) method to estimate the proficiency and the MFI criterion to select the next item. At this stage, test security can be controlled using the appropriate test algorithms, and, in the post-test, collateral information in the form of the pretest proficiency estimate of a student can be used as the empirical prior to estimate the proficiency of the respondent (see below for description).

(iii) *Stopping step*.—This is the step where the test checks that a certain criterion has been met and the test ends. We chose length to be the stopping criterion, such that the FCI-CAT stops after a predetermined number of items have been administered, ranging from 1 to 30. As we mentioned above, we considered also cases in which the pretest length and the post-test length differ.

(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. One can use collateral information for the final proficiency estimation in the post-test as in the test step.

In our analysis, we implemented *content balancing* for the FCI-CAT as in our previous study [64]. Content balancing is ensuring that 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 [36] (for example, to be the same as what is found in the FCI itself). Doing so ensures that items from each subgroup are administered. Various algorithms exist to control for content balancing, but the CATR package [36] allows use only of the simplest option, the *constrained content balancing method* [45], and so we chose this method.

C. Test security

1. Metrics

The risks of item exposure are classified into two categories: test-retest exposure and peer-to-peer exposure [45,65]. Test-retest exposure is the risk that respondents memorize the items of the pretest, and that they utilize this knowledge on the post-test. Peer-to-peer exposure is the risk that respondents share the items of a test with future respondents. Different statistics are used to evaluate each of these risks. Chang and Asley [65] calculated the magnitude of the risk from test-retest exposure by averaging the percentage of items that overlapped between two CAT-based administrations of the same assessment given back-to-back to respondents of a given proficiency. Such an approach can be used to measure the risk over a very short time span. We, on the other hand, are interested in evaluating the risk of test-retest memorization between administrations of the FCI where learning may have occurred between the two administrations. Therefore, we evaluated the risk of test-retest memorization by averaging the proportion of items administered both pre- and post-test to a given respondent with different pre- and post-proficiency estimates. Our metric differs from that of Chan and Asley, because they calculated the rate of overlapped items on subsequent administrations when proficiency estimate for a given respondent is kept fixed. The respondents in our study, on the other hand, generally have different pre- and post-proficiency estimates. In order to distinguish our metric from that of Chan and Asley, we call our metric as *pre-post overlap rate*. Specifically, we calculate the pre-post overlap rate by dividing the average number of overlapping items by the total number of items administered on the post-test.

The risk of sharing the items between respondents is evaluated by the *peer-to-peer overlap rate* (or simply, the *overlap rate* as in Ref. [52]), which is defined as the average proportion of items that are shared by two randomly selected respondents. For both of these statistics (pre-post overlap rate and peer-to-peer overlap rate), higher test security is indicated by lower values.

As an example of the calculation of these statistics, let us consider the two half-length FCI (HFCI) assessments of Han *et al.* [25]. Since their two HFCI assessments both have 14 items, and 3 of those items overlap, the pre-post overlap rate is calculated as 21.4%. Since their HFCIs are linear, the peer-to-peer overlap rate is 100% for both assessments. We will compare these values to our results in the following analysis. Note that our analysis is different from that of Han *et al.* in many ways: using Monte Carlo simulation, calculating Cohen's d, analyzing the class size dependence, and so on. As such, we do not aim to conclude whether our method is more secure and efficient than that of Han *et al.*; rather, we use the results of Han *et al.* as a benchmark to clarify the advantages and issues of the FCI-CAT.

2. Algorithms

We utilized algorithms to enhance test security when the item to be administered is selected, both in the initial step and in the test step. Out of the various methods to control test security, we utilized the two algorithms considered by Barrada *et al.* [52] which are available in the package CATR: the progressive (PG) method and the proportional (PP) method. Similar to Barrada *et al.*, we compared these two algorithms with the maximum likelihood weighted information (MLWI) criterion, the posterior Kullback-Leibler (KLP) criterion, and the MFI criterion to examine effects on test efficiency. We found that there is no significant difference between the results of MLWI, KLP, and MFI; therefore, in this paper, we describe the results of PG, PP, and MFI.

The progressive (PG) method [66,67] is a method to decrease item overexposure at the early stages of CAT. The PG method, like the MFI criterion, chooses as the next item the one which maximizes the *objective function*. In the case of MFI, this objective function is just the Fisher information function. In the case of the PG method, the objective function for the jth item is the weighted sum of a random component R_j and an information component I_j based on the Fisher information, in the form $(1-W)R_j + WI_j$, where W is the weight function. At the beginning of the test, W is close to zero and the random component dominates: the selection of items is close to random. As the number of items administered increases, W increases to one and the information component dominates: the selection of items closely resembles the MFI criterion. The increase of the weight function is controlled by the *acceleration parameter*, which marks the speed at which the weight of the random component is reduced and, thus, the speed at which the importance of item information increases. We selected the value of the acceleration parameter as 1, since literature has described this value as resulting in marked improvement in security with minimal detriment to accuracy [52].

The proportional method [67,68] is also a method to decrease item overexposure at the early stages of CAT. In this method, an objective function is not used to order the items; instead, the probability of selecting a given item is calculated as a function of the Fisher information function. Specifically, P_j, the probability of selecting item j, is determined by the Fisher information I_j raised to a given power H, in the form $P_j \propto I_j^H$, where H is equal to 0 at the beginning of the test and increases as the test advances. This means that the test starts with completely random selection (P_j is same for all items) and becomes similar to MFI at the end of the test. The increase of the power is controlled by the acceleration parameter which plays a similar role as in the PG method. We selected the value of the PG acceleration parameter to also be 1, as it is the default value of the package CATR.

D. Test efficiency

Generally, introducing an algorithm to enhance test security comes at a cost to test efficiency, resulting in longer test lengths. We can more than compensate for this, however, by utilizing collateral information pertaining to the respondents' proficiency level obtained before the test administration.

1. Metrics

As we described above, test efficiency is defined by the minimal test length yielding a desired accuracy and precision of the measurement. To calculate the accuracy and precision, the metric we use is the standardized mean difference, Cohen's d in particular, as in our previous study [34].

The population parameter of Cohen's d is given by [43]

$$d = \frac{\mu_{\mathrm{post}} - \mu_{\mathrm{pre}}}{\sigma}, \tag{2}$$

where μ_{pre} and μ_{post} are the population means for the pretest and post-test, respectively, and σ is the standard deviation of either pre- or postpopulation (we assume that the two population standard deviations are the same, as is done in most parametric data analysis techniques [43]).

We represent the pair of the pre- and post-test lengths as the vector $\boldsymbol{l} = (l_{\mathrm{pre}}, l_{\mathrm{post}})$, and express the estimator for d of the test length $\boldsymbol{l}$ as $\hat{d}_{\boldsymbol{l}}$. From within the family of estimators for d, we use the following definition for repeated measures [43],

$$\hat{d}_{\boldsymbol{l}} = \frac{\bar{\theta}^{\boldsymbol{l}}_{\mathrm{post}} - \bar{\theta}^{\boldsymbol{l}}_{\mathrm{pre}}}{s_{\boldsymbol{l}}}, \tag{3}$$

where $\bar{\theta}^{\boldsymbol{l}}_{\mathrm{pre}}$ and $\bar{\theta}^{\boldsymbol{l}}_{\mathrm{post}}$ are the means of the final estimated proficiencies of the $\boldsymbol{l}$-length pre- and post-test, respectively. [The superscript $\boldsymbol{l}$ takes l_{pre} (l_{post}) for the subscript pre (post).] $s_{\boldsymbol{l}}$ is the pooled standard deviation for dependent (paired) data defined as,

$$s_{\boldsymbol{l}}^2 = \frac{(s^{\boldsymbol{l}}_{\mathrm{pre}})^2 + (s^{\boldsymbol{l}}_{\mathrm{post}})^2 - 2r_{\boldsymbol{l}} s^{\boldsymbol{l}}_{\mathrm{pre}} s^{\boldsymbol{l}}_{\mathrm{post}}}{2(1-r_{\boldsymbol{l}})}, \tag{4}$$

where $s^{\boldsymbol{l}}_{\mathrm{pre}}$ and $s^{\boldsymbol{l}}_{\mathrm{post}}$ are the standard deviations of the final estimated proficiencies of the $\boldsymbol{l}$-length pre- and post-test, respectively, and $r_{\boldsymbol{l}}$ is the Pearson correlation coefficient.

We represent the accuracy and precision by the bias and standard error (the correspondence is reciprocal, respectively). Then, we summarize these measures in terms of the root-mean-square error (RMSE). The RMSE at test length $\boldsymbol{l}$ is defined by the following equation [69], which equals the square root of the sum of the squared bias and squared standard error,

$$\mathrm{RMSE}_{\boldsymbol{l}} = \sqrt{E[(\hat{d}_{\boldsymbol{l}} - d)^2]} = \sqrt{B_{\boldsymbol{l}}^2 + \mathrm{SE}_{\boldsymbol{l}}^2}, \tag{5}$$

where $E(x)$ is the expected value of x. The bias $B_{\boldsymbol{l}}$ is defined by,

$$B_{\boldsymbol{l}} = E(\hat{d}_{\boldsymbol{l}}) - d. \tag{6}$$

The standard error $\mathrm{SE}_{\boldsymbol{l}}$ is given by

$$\mathrm{SE}_{\boldsymbol{l}} = \sqrt{E\{[\hat{d}_{\boldsymbol{l}} - E(\hat{d}_{\boldsymbol{l}})]^2\}}. \tag{7}$$

As the desired accuracy and precision, we used the value of $\mathrm{RMSE}_{\boldsymbol{l}}$, which is obtained when the full-length FCI is administered as a paper-and-pencil test: the test length of the pre- and post-test are both 30 and the collateral information (CI) is not used. That is, we define test efficiency by the minimal value of the sum of the elements of the vector $\boldsymbol{l}$, which satisfies

$$\mathrm{RMSE}_{\boldsymbol{l}}^{\text{with CI}} \leq \mathrm{RMSE}_{(l_{\text{pre}}=l_{\text{post}}=30)}^{\text{without CI}}. \tag{8}$$

In our Monte Carlo study (described below), we generated pre- and postresponses to the FCI-CAT and calculated $\hat{d}_{\boldsymbol{l}}$ 10 000 times for each $\boldsymbol{l}$ to analyze the sampling distribution of $\hat{d}_{\boldsymbol{l}}$. For example, $E(\hat{d}_{\boldsymbol{l}})$ in Eq. (6) is estimated by taking the average of 10 000 samples of $\hat{d}_{\boldsymbol{l}}$.

2. *Algorithms*

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.

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 simulation selects the item with the largest Fisher information for the current estimate of the respondent's proficiency. At the beginning of the pretest, 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 [36], as we described above. At the beginning of the post-test, when calculating the Fisher information, we can utilize the pretest proficiency estimate of a given student as collateral information to improve the test efficiency for that student. Generally, the farther the initial proficiency estimate is from true proficiency of the respondent, the slower the algorithm converges [53]. To decrease this gap, there are two methods utilizing collateral information. One possibility is to directly use a given student's proficiency estimate from the pretest as the initial proficiency estimate of the post-test [55]. The other possibility is predicting the initial proficiency estimate of the post-test by a regression model fit to previous paired pre- and post-proficiency estimates [53]. In principle, this second method could be used if much pre- and post-test data had formerly been collected for a given instructional approach for a given instructor. The FCI, however, is generally used to measure the effectiveness of new teaching approaches, and so these data would not be available. Therefore, we use the first approach, namely, directly using the pretest proficiency estimate of each respondent to calculate the Fisher information and thereby determine the first item on the post-test.

Test step.—CAT selects the next item based upon the estimated proficiency level of the respondent at that point in the assessment. We estimated the proficiency level using the EAP estimator, which is based upon 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 [36]. We consider the model with 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 (assuming a normal distribution is common in proficiency estimation using Bayesian item response modeling [36]). On the pretest, 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, $g(\theta^i) \sim \mathcal{N}(0, 1)$ [36]. On the post-test, we can utilize the pretest proficiency estimate of an ith respondent, $\hat{\theta}^i_{\text{pre}}$ as collateral information for the prior distribution. Specifically, we chose the prior distribution as the normal distribution, with $\mu^i = \hat{\theta}^i_{\text{pre}}$. If, in the initial step, a regression is made between pre- and post-proficiency estimates (method 2 above), then σ^i can be estimated [53]. However, in the model directly using the proficiency estimate (method 1 above), 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) [35]. Therefore, as in Ref. [54], we set $\sigma^i = 1$, which is the same value used when no collateral information is available.

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

E. Procedure of simulation study

To analyze the security and efficiency of the FCI-CAT, we conducted two simulations that are commonly used in

CAT development, a Monte Carlo simulation and a *post hoc* simulation [46]. Monte Carlo simulations generate responses with pseudorandom numbers, while *post hoc* simulations utilize empirical data. To ensure that the simulated data are compatible with actual data one might obtain from a real classroom, we examined the consistency of the results of these simulations.

1. Monte Carlo study

In this simulation, we followed a two-step process to generate paired pre- and post-responses similar to our previous study [34]. In the first step, we generated a pair of true proficiencies for a given simulee, one corresponding to the pretest and one corresponding to the post-test. For both pre- and post-tests, true proficiencies were generated from the bivariate normal population distributions with designated population parameters. We chose these parameters such that the estimates by the simulation for the 30-item length test are as close as possible to the statistics calculated with our empirical data previously obtained (for details of the empirical data, see the description of dataset β in Ref. [34]). These parameters were pretest true proficiency mean $= 0.44$, post-test true proficiency mean $= 0.75$, standard deviation for both sets of true proficiency $= 0.82$, and correlation $= 0.98$. From these parameters, we generated a pair of pre- and post-true proficiencies for each of 100 000 simulees.

In the second step, we generated the responses for the FCI-CAT. As discussed above, the EAP method was used to estimate the proficiency for the respondent, the item selection method (e.g., the MFI criterion) was then used to choose the next item based upon that estimated proficiency, and the process repeated until reaching the predetermined test length. This was done for the simulee both on the pretest and on the post-test. In this manner, we generated paired pre- and postresponses and estimated proficiencies for 100 000 simulees for each vector $\boldsymbol{l}$ of the FCI-CAT.

Finally, for each vector $\boldsymbol{l}$ of the FCI-CAT, from the 100 000 paired pre- and postresponses, we resampled with replacement, 10 000 paired responses for each simulee in various class sizes (40, 60, 80, 100). For example, in the case with class size of 100, we resampled 10 000 times 100 paired responses with replacement from the 100 000 paired responses. Then, we calculated the estimate $\hat{d}_{\boldsymbol{l}}$ and the corresponding $\mathrm{RMSE}_{\boldsymbol{l}}$.

2. *post hoc* study

Since the responses generated via the Monte Carlo simulation are just imaginary responses, we conducted another simulation using empirical responses (that is, a "*post hoc* simulation"). *post hoc* simulations are commonly used to determine how short a CAT-based assessment can be without excessively sacrificing accuracy and precision [46,60]. In a *post hoc* simulation, a CAT-based assessment is simulated for each respondent based upon their actual responses to the full-length assessment. For example, if the CAT simulation for a given respondent determines that the respondent should next be administered item 8, the simulation algorithm would look up and utilize the actual answer of the respondent to item 8. In this way, although examinees have not taken the FCI-CAT, we can simulate their testing experience as if they had. In the *post hoc* simulation, we used the same empirical responses as our above Monte Carlo study (dataset β) and examined the consistency of the results.

We confirmed that the results of the Monte Carlo study and the *post hoc* simulation were consistent. In particular, all of the absolute differences between the estimates in the *post hoc* simulation and the expected values $E(\hat{d}_l)$ calculated with the Monte Carlo simulation are less than the standard errors of $E(\hat{d}_l)$, with the exception of the extremely short post-test length ($l_{\mathrm{post}} = 1$). (Details of this comparison are reported in Ref. [34] for the specific case of using the MFI criterion without collateral information or content balancing.) This consistency supports the adequacy of our Monte Carlo study and its relevance in terms of administering the FCI-CAT in real classrooms.

III. RESULTS

A. Effects of test security algorithms on RMSE

Figure 2 shows the relationship between the test length and the RMSE of Cohen's d when test security algorithms are implemented with the FCI-CAT. The test length of the pre- and post-test is fixed to be the same and the class size is set to 40. [We begin the discussion of our results with a class size of 40, as (i) it is a typical lower bound for the university classes from which our empirical data originates and (ii) it is large enough to stay clear of small sample effects.] We also compared the effect of content balancing. The plus symbol shows the result with content balancing (CB) and the circle symbol shows the result without content balancing. For the following explanation, we represent the test length as $l(= l_{\mathrm{pre}} = l_{\mathrm{post}})$. If only one item is administered ($l = 1$), the item is selected randomly in the PG and PP methods, thereby the values of RMSE of both methods are almost the same as each other and much larger than that of the MFI criterion. If additional items are administered ($l \geq 2$), the RMSE of the PP method is smaller than that of the PG method. This result indicates that the PP method switches from selecting items at random to selecting items based on the information function at an earlier point than the PG method. This is further illustrated by the fact that the gap between with and without content balancing appears earlier for the PP method (In the PG method, the gap appears when $l \geq 9$ but in the PP method, the gap appears when $l \geq 3$). Since content balancing interferes with selecting informative items in the MFI criterion in the early stage [64] (which can be seen from the red curve in

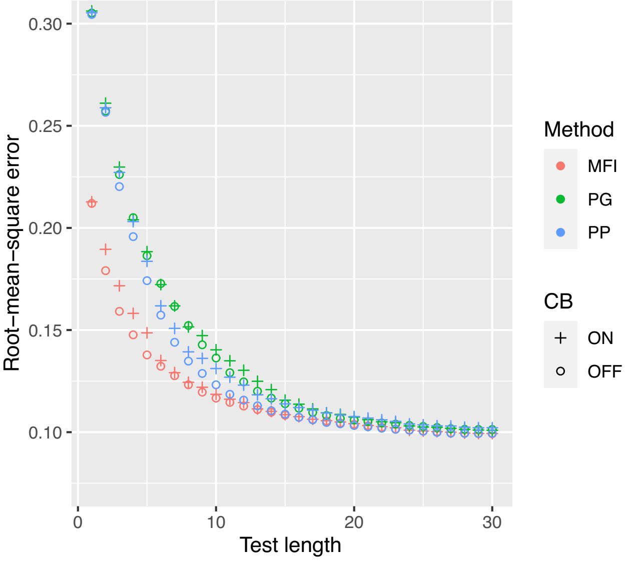

FIG. 2. Comparison of the effects of test security algorithms on RMSE (maximum Fisher information, red; progressive, green; proportional, blue). The class size is 40 and the test length of the pre- and post-test is fixed to be the same. The plus symbol shows the result with content balancing and the circle symbol shows the result without content balancing.

Fig. 2), appearance of the gap between with and without content balancing in the green and blue curves indicates that it switches from selecting items at random to selecting items based on the information function. These results may vary with the value of the acceleration parameter, which we set to be 1 in both methods as we described above. Finally, at $l = 30$ the RMSE in both test security methods reaches that of the MFI criterion. Although we described the results of the case when the class size is 40 students, we found the results to be similar for the cases of class size equal to 60, 80, and 100.

B. Effects of test efficiency algorithms on RMSE

Figure 3 shows the effects of introducing collateral information (CI) into each test security algorithm when content is balanced and the class size is $N = 40$. This figure demonstrates the possibility of allowing the pretest and post-test to vary in length. Each graph in Fig. 3 shows how RMSE depends upon pretest length and post-test length. In the graphs, the light (dark) color indicates a low (high) value of RMSE with a bin width of 0.02 (except for the highest bin, where RMSE > 0.26). For comparison with MFI with collateral information (upper right), a plot of MFI without collateral information (upper left) is also included. Overall, we can see that RMSE has decreased with the introduction of collateral information. In particular, the decrease is large in the region where the post-test length is small (the lower-right corner of the upper-right graph is brighter than that in the upper-left graph). This result follows from the fact that the post-test initial proficiency estimate was, in most cases, closer to the true proficiency, as a result of using collateral information. A more accurate starting point results in a smaller bias and hence smaller RMSE.

In Fig. 3, we can compare the two test security algorithms: progressive (lower-left) and proportional (lower-right) methods with MFI (upper-right) when collateral information is introduced. Overall, we can see that the RMSE of PG and PP methods are larger than that of MFI. The increase is particularly large when the pretest length or post-test length is small. This is because in the PG method and PP method the items are selected somewhat randomly for small test lengths, as we described above.

As shown in the graphs, in all three methods, if the asymmetry of the pre- and post-test lengths is large, RMSE becomes large. In other words, RMSE is small if both tests are long or if both tests are short. This trend is explained as follows. The bias of Cohen's d can be written as $B = B_{\mathrm{post}} - B_{\mathrm{pre}}$, where B_{post} is the bias of the means of the estimated proficiencies of the post-test divided by the standard deviation and B_{pre} is the term calculated for the pretest in the same way. We found that both B_{post} and B_{pre} are negative for any test length and become smaller (larger in magnitude) when the test length becomes smaller. For a post-test length much smaller than a pretest length, the magnitude of B_{post} is much larger than that of B_{pre}. (Similarly, for a pretest length much smaller than a post-test length, the magnitude of B_{pre} is much larger than that of B_{post}). In this case, the magnitude of B and RMSE becomes large. When the pre- and post-test lengths are close to each other, the values of B_{post} and B_{pre} are also close to each other and the difference, B, becomes small. The dependence of the standard error on the test length is sufficiently small to have a negligible impact on this effect.

C. Calculating test efficiency of the FCI-CAT

Following the definition of test efficiency in Eq. (8), we calculated the minimal test length of the FCI-CAT yielding the RMSE as what is obtained when the full-length FCI is administered as a paper-and-pencil test: the test length of the pre- and post-test are both 30 and the collateral information is not used. In this case, the RMSE is calculated as 0.10 for a class size of 40 (see Fig. 2). Having determined this reference value, we find the $\boldsymbol{l}$ vector ($\boldsymbol{l}_{\mathrm{min}}$) for which the RMSE is less than that value (0.10 for $N = 40$) and the total number of items is minimized. For PP with $N = 40$, this approach results in several possible combinations of l_{pre} and l_{post} with the same minimal values of l_{total}. In such a case, we chose the combination which minimizes the RMSE. The result for each item selection method for a class size of 40 was $\boldsymbol{l}_{\mathrm{min}} = (16, 13)$ in MFI, $\boldsymbol{l}_{\mathrm{min}} = (21, 15)$ in PG, and $\boldsymbol{l}_{\mathrm{min}} = (17, 16)$ in PP.

We can interpret these results with the help of Fig. 3. The total number of items is given by l_{total} in the expression

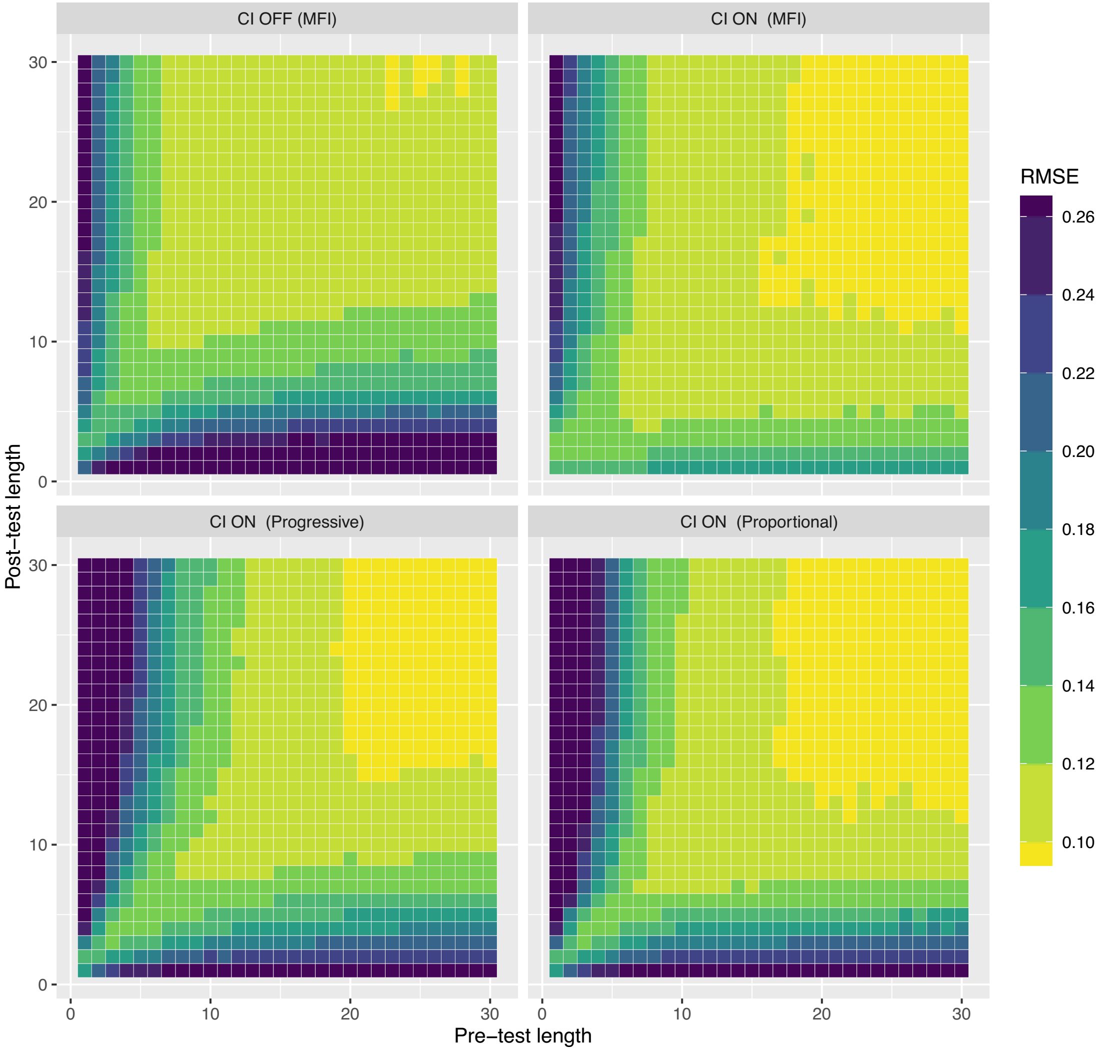

FIG. 3. Comparison of the effects of collateral information on RMSE for each test security algorithm (maximum Fisher information, progressive, and proportional). The x (y) axis is the pre- (post-) test length. The class size is 40. The results show the case with content balancing. The light (dark) color indicates a low (high) value of RMSE, the bin width of the RMSE is set to 0.02 (except for the highest bin, where RMSE > 0.26).

$l_{\text{total}} = l_{\text{pre}} + l_{\text{post}}$. This expression can be represented as a line in the graphs above: $l_{\text{post}} = l_{\text{total}} - l_{\text{pre}}$, which has a slope of -1 and y intercept of l_{total}. We increase the y intercept (l_{total}) until the line first passes through the area RMSE < 0.1 in Fig. 3. For example, in MFI with collateral information, we found that the line first passes through the area when $l_{\text{total}} = 29$. As in Fig. 4, this line $l_{\text{post}} = 29 - l_{\text{pre}}$ passes through the area RMSE < 0.1 at $l_{\text{pre}} = 16$ and $l_{\text{post}} = 13$, which gives the minimal test length.

For the other class sizes, the reference values of RMSE were somewhat different than with $N = 40$ (the reference values are 0.083, 0.074, and 0.067 for class size of 60, 80, and 100, respectively), but $\boldsymbol{l}_{\text{min}}$ was otherwise calculated in the same manner. The results are shown in Table I.

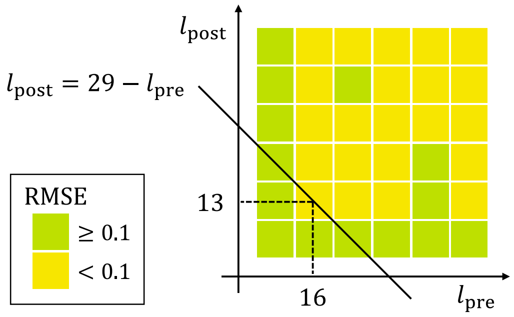

FIG. 4. The line $l_{\text{post}} = 29 - l_{\text{pre}}$ passing through the area RMSE < 0.1 in MFI with CI (Fig. 3, upper right). This derives the minimal test length as $l_{\text{pre}} = 16$ and $l_{\text{post}} = 13$ ($l_{\text{total}} = 29$).

TABLE I. Dependence of minimal test length (l_{pre}, l_{post}) on class size.

Class size (N)	MFI	PG	PP
40	(16, 13)	(21, 15)	(17, 16)
60	(13, 10)	(17, 15)	(15, 13)
80	(13, 9)	(15, 14)	(15, 11)
100	(11, 9)	(15, 12)	(12, 10)

Note that in all of these combinations, the post-test length is less than that of the pretest. This trend is explained as follows. As we explained above, RMSE is minimized when $B = B_{\mathrm{post}} - B_{\mathrm{pre}} \sim 0$. The magnitude of $B_{\mathrm{pre}}(B_{\mathrm{post}})$ decreases when the pre (post-) test length increases. For a given pretest length, B_{post} is able to equal B_{pre} with a smaller post-test length due to the use of collateral information. Note also that as the class size is increased, the minimal test length decreases. This is because the RMSE converges (comes close to the RMSE value for $l = 30$) faster for larger class sizes, as we found in Ref. [34].

As we described above, we found that the accuracy and precision of the full-length FCI can be attained with a smaller test length when FCI-CAT is used. This finding resolves an issue of our previous study [34]. In the previous study, we examined the minimal test length of the FCI-CAT without "excessively" compromising accuracy and precision; however, since the criterion of the "excessiveness" (say, 5% or 10%) is arbitrary, we could not specify a definite minimal test length. In the present study, there is no such arbitrariness: we have clearly defined the minimal test length for a given class size and item selecting algorithm.

D. Evaluation of test security of the FCI-CAT

We evaluated the risk of test-retest exposure by the pre-post overlap rate, and we evaluated the risk of peer-to-peer exposure by the peer-to-peer overlap rate. As we described above, for both of these statistics, higher test security is indicated by lower values. Table II shows the results of calculating the pre-post overlap rate and the peer-to-peer overlap rate for each item selection method. In the top and middle of the table, we present the results of $N = 40$ without and with collateral information. In the bottom of the table, we present the rates at the minimal test lengths ($\boldsymbol{l}_{\min}$) shown in Table I for each class size ($N = 40$, 60, 80, 100). Five major observations on test security can be made from the table. First, both the pre-post overlap rate and the peer-to-peer overlap rate of the PG method are smallest for a given $\boldsymbol{l}$: the PG method is the most secure of the three methods for a given test length. This is a natural consequence, since, as discussed above, the PG method is closer to random selection than the other methods. Second, the

TABLE II. Pre-post overlap rate and peer-to-peer overlap rate of various test lengths for three item selection methods. In the top and middle of the table, we present the results of $N = 40$ without and with collateral information. In the bottom of the table, we present the rates at the minimal test lengths ($l_{\min}$) shown in Table I for each class size ($N = 40$, 60, 80, 100).

			MFI		PG		PP	
	l_{pre}	l_{post}	Peer-to-peer	Pre-post	Peer-to-peer	Pre-post	Peer-to-peer	Pre-post
CI OFF	10	10	67.9	83.4	34.9	35.2	43.6	47.4
	10	15	67.9	60.2	34.9	34.9	43.6	44.2
	10	20	67.9	47.9	34.9	34.7	43.6	40.3
	15	10	77.9	90.0	55.3	52.5	65.5	66.7
	15	15	77.9	87.1	55.3	56.8	65.5	69.7
	15	20	77.9	70.6	55.3	55.7	65.5	63.0
	20	10	86.0	96.0	74.1	69.6	79.5	81.1
	20	15	86.0	94.1	74.1	74.4	79.5	84.3
	20	20	86.0	92.0	74.1	76.1	79.5	83.1
CI ON	10	10	67.9	77.7	34.9	35.1	43.5	47.0
	10	15	67.9	57.9	34.9	34.8	43.5	43.8
	10	20	67.9	46.7	34.9	34.6	43.5	40.1
	15	10	77.9	89.0	55.3	52.5	65.3	66.8
	15	15	77.9	84.5	55.3	56.7	65.3	69.3
	15	20	77.9	69.4	55.3	55.6	65.3	62.7
	20	10	86.0	95.5	74.1	69.8	79.4	81.6
	20	15	86.0	94.0	74.1	74.5	79.4	84.3
	20	20	86.0	90.6	74.1	76.0	79.4	82.8
CI ON	$l_{\min}(N = 40)$		82.2	88.5	74.1	74.5	71.4	75.1
	$l_{\min}(N = 60)$		70.9	84.3	63.4	63.8	65.4	69.2
	$l_{\min}(N = 80)$		72.2	83.0	55.3	55.9	60.6	63.7
	$l_{\min}(N = 100)$		67.9	81.5	55.3	53.7	52.4	54.6

pre-post overlap rate and the peer-to-peer overlap rate increase with pretest length. This result indicates that, from the point of view of test security, it is preferable that the pretest length be smaller. Third, when collateral information is used, the pre-post overlap rate slightly decreases for the MFI algorithm. However, it hardly changes for the PG and PP methods because the items are selected almost randomly in the early stage of these methods already, and so the personalized entry point obtained with the use of collateral information is redundant. Fourth, at the minimal test length, the pre-post overlap rate and the peer-to-peer overlap rate of the PG and PP methods take similar values. Since the minimal test length is less for the PP method (see Table I), it may be preferable to use the PP method when considering both test efficiency and test security. Fifth, for all of the item selection methods at the minimal test length, the pre-post overlap rate is more than two times larger than that of the HFCI (21.4%) and the peer-to-peer overlap rate is much smaller than that of the HFCI (100%). As we expected, the FCI-CAT mitigates the risk of the peer-to-peer exposure compared to the HFCI. On the other hand, even when an algorithm for test security is utilized, the risk from test-retest exposure of the FCI-CAT is much larger than that of the HFCI.

Our results in Sec. III C are that if one intends to maximize test efficiency, the pretest length should be slightly larger than the post-test length. On the other hand, our results in this section are that if one intends to maximize test security, the pretest length should be smaller and the post-test length should be larger. These results illustrate a tradeoff that must be considered when determining the test length in the administration of the FCI-CAT. Overall, if one desires a balance of these two goals, it would be reasonable to choose equal pretest and post-test lengths.

IV. DISCUSSION

A. Summary

We improved a computerized adaptive testing (CAT)-based version of the FCI considering test security and test efficiency. First, we implemented algorithms for test security to reduce item overexposure. Second, we improved test efficiency by utilizing the pretest proficiency estimate of each respondent for selecting items and estimating respondent proficiency level in the post-test. To shorten the test length further, we also examined the case when the test length is different for the pre- and post-test. We conducted a Monte Carlo simulation to analyze how implementing these algorithms affects the accuracy and precision of Cohen's d and calculated the minimal test length of the FCI-CAT whose accuracy and precision are equivalent to that of the full-length FCI. Consequently, we obtained the following three findings: (i) By using collateral information, the accuracy and precision of the full-length FCI can be achieved with fewer items via the FCI-CAT. (ii) For a class size of 40, we can *also* control for test security with pre- and post-test lengths of the FCI-CAT totaling 33 items (17 items on the pretest and 16 items on the post-test), thereby reducing the testing time to 55%. (iii) If one's goal is to maximize test efficiency, the pretest length should be slightly larger than the post-test length. On the other hand, if the goal is to maximize test security, the pretest length should be smaller and the post-test length should be larger. If one desires a balance of these two goals, it would be reasonable to choose equal pretest and post-test lengths.

The results of the Monte Carlo study and the *post hoc* study were consistent, which supports the adequacy of our Monte Carlo study and its relevance in terms of conducting the FCI-CAT in real classrooms.

B. Limitations and future work

1. Population dependence

It is important to note that our findings stem from simulations which were specific to a given value of true Cohen's d determined from our empirical data. Specifically, in our empirical data, as discussed in Sec. II E, the pretest true proficiency mean $= 0.44$ and post-test true proficiency mean $= 0.75$, both positive values. In cases where the pretest mean is negative and the post-test mean is positive, it remains an open question of whether collateral information actually improves the test efficiency (in comparison to the default of assuming a prior population mean proficiency value of zero). Furthermore, our data consist exclusively of responses from Japanese students. The estimates for the item parameters of the FCI vary, depending on the students taking the FCI. Item parameters might be different for students in different countries, might be gender dependent, might be different for students in calculus-based physics courses vs algebra-based courses, and so on. Additional research is necessary to see how results are different for other student populations. Concretely, for a given population, empirical data from the FCI should be accumulated to investigate the typical range of the mean, standard deviation, and correlation of the pre- and post-tests. These values should then be used to conduct simulation studies with that population that are similar to what we have presented in this paper.

2. Improving test security

As we described in Sec. III D, we compared test security of the FCI-CAT to that of the HFCI and found that for all of the item selection methods at the minimal test length, the pre-post overlap rate is three times larger than that of the HFCI. In order to reduce the pre-post overlap rate, one could use an algorithm in the post-test to administer the items which were not administered in the pretest as in Ref. [45], though it would compromise test efficiency. One of the other approaches to reduce the pre-post overlap rate is to enlarge the item pool to choose from, namely, to add new items to the FCI. To do this, for example, one could include the items of other versions of the FCI [70–73] to the item pool of the FCI-CAT or one could create new items following the original approach. Doing so would require the process of validating the items and then equating them to the original items.

3. Validating the FCI-CAT

The FCI-CAT can be validated and improved in means other than test security as well. Regarding validity of the computer based test form itself, Nissen *et al.* [74] showed that student performance on the FCI is equivalent for the online linear CBT (computer-based test administered out-of-class, non-CAT) test form and for the paper-and-pencil (in-class) test form. Future work should attend to showing that the FCI-CBT and FCI-CAT are measuring the same constructs. Then, by a "chain of validity," we can expect the FCI-CAT and the paper-and-pencil administrations to also be measuring the same constructs. One of the differences of the FCI-CBT and FCI-CAT is ordering of the questions. In IRT, the effects of item ordering is examined by evaluating local independence. We found that our FCI dataset has sufficient local independence at the whole test level [34]; therefore, we expect the effect of item ordering to not be large. However, it is meaningful to confirm the validity of the CAT test form by administering the FCI-CAT in real classes and comparing the result to that of the FCI-CBT. It would also be helpful to examine the reliability of the FCI-CAT via test-retest studies in real classes as described in Ref. [39]. This could further assess the unidimensionality assumed in our analysis of the FCI-CAT [75]. If a test is demotivating, students may put less thought into later items than into earlier items. Such an effect is typically indicated by a low test reliability. It is necessary to check that such an effect is not occurring for the FCI-CAT in real classes.

4. How to further improve the FCI administration

In addition to the methods we used in this study, there are many options to explore that may further improve the efficiency of the FCI-CAT. As we described above, the closer the initial proficiency estimate is from the true proficiency, the faster the algorithm converges. Therefore, when we use the pretest proficiency estimate as collateral information for the post-test, if the respondent's true proficiency gain is small, fewer items are necessary to achieve a required accuracy and precision. In our study, we used the length criterion as the stopping step. An alternative is to use the precision criterion as the stopping step. This criterion administers as many items as necessary until a prespecified accuracy and precision are obtained [36]. Improvement might also be found by using other models in IRT (e.g., partial credit grading model [76] and multidimensional models [77–79]). Finally, although we assumed a normal distribution in the algorithm of the post-test proficiency estimation, future work could attend to investigating other prior distributions that correspond more closely to authentic learning progressions. Future work should investigate these alternative criteria and models because it may allow for a further shortening of the test. The FCI-CAT can be improved in other means as well. For example, one may wish to remove gender unfair items [8,80–82] or to drop one item from each locally dependent pair [83]. Another possibility is using multistage testing [36] instead of CAT, which allows respondents to review their item responses within each module [84]. As future work is done to further improve the FCI-CAT, the results we have presented in this paper can serve as reference values for comparison with the results obtained from studies focusing on these other aspects of instrument validity.

5. Deploying the improved FCI-CAT

Concurrently with the analyses we reported here, we have conducted a trial administration of the FCI-CAT. In the deployment of the FCI-CAT, we utilized the Concerto platform [85], which is an open source online adaptive testing platform. Students used their smart phones to take the FCI-CAT, enabling them to take the survey in the classroom instead of moving to a place where there are computers (computer room or their home, etc.). This allows for a greater concentration of students, since instructors can monitor the students during the test. We next plan to administer the FCI-CAT both pre- and postsemester to analyze the effect size distribution. We will compare these results with what we discussed above from our simulations.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant No. JP19K03135.

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

[2] N. S. Rebello, and D. A. Zollman, The effect of distracters on student performance on the Force Concept Inventory, Am. J. Phys. **72**, 116 (2004).

[3] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).

[4] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).

[5] 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).

[6] J.-i. Yasuda and M.-a. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, Phys. Rev. ST Phys. Educ. Res. **9**, 010113 (2013).

[7] 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).

[8] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[9] J.-i. Yasuda, N. Mae, M. M. Hull, and M.-a. Taniguchi, Analyzing false positives of four questions in the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010112 (2018).

[10] P. Eaton, Evidence of measurement invariance across gender for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **17**, 010130 (2021).

[11] Y. Shoji, S. Munejiri, and E. Kaga, Validity of Force Concept Inventory evaluated by students' explanations and confirmation using modified item response curve, Phys. Rev. Phys. Educ. Res. **17**, 020120 (2021).

[12] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[13] J.-i. Yasuda, H. Uematsu, and H. Nitta, Validating a Japanese version of Force Concept Inventory (in Japanese), J. Phys. Educ. Soc. Jpn. **59**, 90 (2011).

[14] G. W. Hitt, A. F. Isakovic, O. Fawwaz, M. S. Bawa'Aneh, N. El-Kork, S. Makkiyil, and I. A. Qattan, Secondary implementation of interactive engagement teaching techniques: Choices and challenges in a Gulf Arab context, Phys. Rev. ST Phys. Educ. Res. **10**, 020123 (2014).

[15] A. L. Rudolph, B. Lamine, M. Joyce, H. Vignolles, and D. Consiglio, Introduction of interactive learning into French university physics classrooms, Phys. Rev. ST Phys. Educ. Res. **10**, 010103 (2014).

[16] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, Am. J. Phys. **65**, 45 (1997).

[17] 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).

[18] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).

[19] L. Ding and M. D. Caballero, Uncovering the hidden meaning of cross-curriculum comparison results on the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **10**, 020125 (2014).

[20] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).

[21] A. Maries and C. Singh, Teaching assistants' performance at identifying common introductory student difficulties in mechanics revealed by the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **12**, 010131 (2016).

[22] J. J. Harlow, D. M. Harrison, and A. Meyertholen, Effective student teams for collaborative learning in an introductory university physics course, Phys. Rev. Phys. Educ. Res. **12**, 010138 (2016).

[23] A. Robinson, J. H. Simonetti, K. Richardson, and M. Wawro, Positive attitudinal shifts and a narrowing gender gap: Do expertlike attitudes correlate to higher learning gains for women in the physics classroom?, Phys. Rev. Phys. Educ. Res. **17**, 010101 (2021).

[24] B. R. Wilcox and S. J. Pollock, Investigating students' behavior and performance in online conceptual assessment, Phys. Rev. Phys. Educ. Res. **15**, 020145 (2019).

[25] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[26] D. MacIsaac, R. P. Cole, D. M. Cole, L. McCullough, and J. Maxka, Standardized testing in physics via the World Wide Web, Electron. J. Sci. Educ. **6**, 1 (2002), https://ejrsme.icrsme.com/article/view/7681.

[27] T. F. Scott and D. Schumayer, Central distractors in Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **14**, 010106 (2018).

[28] B. Van Dusen, Lasso: A new tool to support instructors and researchers, American Physics Society Forum on Education, Fall 2018 Newsletter (2018).

[29] E. Kuo, M. M. Hull, A. Elby, and A. Gupta, Assessing mathematical sensemaking in physics through calculation-concept crossover, Phys. Rev. Phys. Educ. Res. **16**, 020109 (2020).

[30] S. M. Stoen, M. A. McDaniel, R. F. Frey, K. M. Hynes, and M. J. Cahill, Force Concept Inventory: More than just conceptual understanding, Phys. Rev. Phys. Educ. Res. **16**, 010105 (2020).

[31] S. Bonham, Reliability, compliance, and security in web-based course assessments, Phys. Rev. ST Phys. Educ. Res. **4**, 010106 (2008).

[32] M. Jariwala, J.-S. S. White, B. Van Dusen, and E. W. Close, In-class vs. Online Administration of Concept Inventories and Attitudinal Assessments, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA* (AIP, New York, 2016), pp. 176–179.

[33] A. Madsen and S. McKagan, Administering research-based assessments online, PhysPort, Expert Recommen-

dations, https://www.physport.org/recommendations/Entry.cfm?ID=93329 (2020).

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

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

[36] D. Magis, D. Yan, and A. A. von Davier, *Computerized Adaptive and Multistage Testing with R* (Springer, Cham, 2017).

[37] C. N. Mills and M. Steffen, *The GRE Computer Adaptive Test: Operational Issues BT—Computerized Adaptive Testing: Theory and Practice* (Springer, Dordrecht, 2000).

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

[39] 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).

[40] 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).

[41] 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).

[42] International vocabulary of metrology—basic and general concepts and associated terms (VIM), Tech. Rep. (Joint Committee for Guides in Metrology), 2008.

[43] H. Cooper, L. V. Hedges, and J. C. Valentine, *Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. (Russell Sage Foundation, New York, 2009).

[44] C. DeMars, Group differences based on IRT scores: Does the model matter?, Educ. Psychol. Meas. **61**, 60 (2001).

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

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

[47] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).

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

[49] 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).

[50] 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).

[51] 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 Sci. Educ. **53**, 45 (2017).

[52] J. R. Barrada, J. Olea, V. Ponsoda, and F. J. Abad, A method for the comparison of item selection rules in computerized adaptive testing, Appl. Psychol. Meas. **34**, 438 (2010).

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

[54] K. M. Morrison, Impact of composite population priors on computer adaptive test proficiency estimates, Ph.D. thesis, Georgia Institute of Technology, 2017.

[55] Q. Xie, The impact of collateral information on ability estimation in an adaptive test battery, Ph.D. thesis, University of Iowa, 2019.

[56] J.-P. Fox, *Bayesian Item Response Modeling* (Springer, New York, 2010).

[57] R Core Team, R: A Language and Environment for Statistical Computing (2020).

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

[59] 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).

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

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

[62] 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).

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

[64] J.-i. Yasuda and M. M. Hull, Balancing content of computerized adaptive testing for the Force Concept Inventory, in *Proceedings of the 2021 Physics Education Research Conference*, Virtual Conference (AIP, New York, 2021).

[65] S. W. Chang and T. N. Ansley, A comparative study of item exposure control methods in computerized adaptive testing, J. Educ. Measure. **40**, 71 (2003).

[66] J. Revuelta and V. Ponsoda, A comparison of item exposure control methods in computerized adaptive testing, J. Educ. Measure. **35**, 311 (1998).

[67] J. R. Barrada, J. Olea, V. Ponsoda, and F. J. Abad, Incorporating randomness in the Fisher information for improving item-exposure control in CATs, Brit. J. Math. Stat. Psychol. **61**, 493 (2008).

[68] D. O. Segall, A sharing item response theory model for computerized adaptive testing, J. Educ. Behav. Stat. **29**, 439 (2004).

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

[70] L. McCullough and D. E. Meltzer, Differences in male/female response patterns on alternative-format versions of the Force Concept Inventory, *Proceedings of the 2001 Physics Education Research Conference, Rochester, NY* (AIP, New York, 2016).

[71] 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).

[72] S. E. Osborn Popp and J. Jane, Can assessment of student conceptions of force be enhanced through linguistic simplification?, American Educational Research Association 2009 (2009).

[73] 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).

[74] 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).

[75] J. Hattie, Methodology review: Assessing unidimensionality of tests and items, Appl. Psychol. Meas. **9**, 139 (1985).

[76] P. Eaton, K. Johnson, and S. Willoughby, Generating a growth-oriented partial credit grading model for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 020151 (2019).

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

[78] 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).

[79] 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).

[80] R. Henderson, J. Stewart, and A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010131 (2019).

[81] M. Mears, Gender differences in the Force Concept Inventory for different educational levels in the United Kingdom, Phys. Rev. Phys. Educ. Res. **15**, 020135 (2019).

[82] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis, Phys. Rev. Phys. Educ. Res. **15**, 020122 (2019).

[83] C. S. Wallace, T. G. Chambers, and E. E. Prather, Item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set, Phys. Rev. Phys. Educ. Res. **14**, 010149 (2018).

[84] D. Yan, A. A. von Davier, and C. Lewis, *Computerized Multistage Testing: Theory and Applications* (CRC Press, Boca Raton, 2014).

[85] K. Scalise and D. D. Allen, Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform, Brit. J. Math. Stat. Psychol. **68**, 478 (2015).