Editors' Suggestion

Optimizing the length of computerized adaptive testing for the Force Concept Inventory

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

Naohiro Mae
Faculty of Engineering Science, Kansai University, Suita, Osaka 564-8680, Japan

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

Masa-aki Taniguchi
Center for Teacher Education, Meijo University, Nagoya, Aichi 468-8502, Japan

(Received 29 September 2020; accepted 26 January 2021; published 9 March 2021)

As a method to shorten the test time of the Force Concept Inventory (FCI), we suggest the use of computerized adaptive testing (CAT). CAT is the process of administering a test on a computer, with items (i.e., questions) selected based upon the responses of the examinee to prior items. In so doing, the test length can be significantly shortened. As a step to develop a CAT-based version of the FCI (FCI-CAT), we examined the optimal test length of the FCI-CAT such that accuracy and precision [which we measure in terms of bias, standard error, and root-mean-square error (RMSE)] of Cohen's d would be comparable to that of the full FCI for a given class size. First, we estimated the item parameters of the FCI items based on the three-parameter logistic model of item response theory, which are used in the algorithm of CAT. For this estimation, we used 2882 responses of Japanese university students. Second, we conducted a Monte Carlo simulation to analyze how the bias, standard error, and RMSE of Cohen's d depend upon the test length. Third, we conducted a *post hoc* simulation to examine the consistency of the Monte Carlo results with what would have been obtained using empirical responses. For this comparison, we used 86 pairs of pre- and post- test responses of Japanese university students. As a result, we found that for a class size of 40, we may reduce the test length of the FCI-CAT to 15–19 items, thereby reducing the test time of the FCI to 50%–63%, with an accompanying decrease in accuracy and precision of only 5%–10%. The results of the Monte Carlo study and the *post hoc* simulation were consistent, which supports the adequacy of our Monte Carlo study and its relevance in terms of administering the FCI-CAT in real classrooms.

DOI: 10.1103/PhysRevPhysEducRes.17.010115

I. INTRODUCTION

The Force Concept Inventory (FCI) is one of the most widely used research-based assessments in physics education [1,2]. It 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 to 30 min to complete the test. The items use everyday speech in order to better elicit what the student personally considers to be correct as opposed to an answer memorized by rote from physics class [3]. Futhermore, the distractors are designed based upon knowledge of students' common naïve conceptions [4,5]. The FCI has been examined from various viewpoints and validated in various regards [6–19], and it has played an important role in analyzing the effects of newly developed pedagogy [20–23].

When administering the FCI in a classroom, instructors typically require about 40 min, including the time needed to orient students to the survey. To administer the survey as a pretest, an instructor must thus be willing to sacrifice nearly an entire class period; a second class period is required if the survey is also given post-test. (The situation is described for the United States in Refs. [24–27], and Japan is similar.) Many instructors feel pressure to cover as much content as possible by the end of the semester, and they are likely to be reluctant to find time in their crowded schedules to administer the assessment.

To avoid using class time for assessments, some instructors administer the assessment via online platforms which

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

enables students to complete the assessment outside of class [24,25]. 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 ungraded conceptual questions online outside of class can decrease response rate and compromise test security [28]. To shorten the test time, Han *et al.* [26,27] divided the FCI into two half-length tests which contain different subsets of the original FCI, but still cover the same set of concepts. They showed, using empirical data (that is, actual data collected by respondents), that the difference of the average normalized gain between either half-length FCI and the full FCI is on the order of 0.03.

We have the same goal in mind as the above-mentioned approaches; namely, our objective is developing a method to reduce the time needed to administer research-based assessments, specifically, in this paper, the FCI. In this study, we suggest an alternative approach to accomplish this goal: the use of computerized adaptive testing (CAT) [29,30], which is the practice of using a computer to administer a set of items taken from an item pool, with the items chosen based on the student's responses to prior items. In one 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 (Fig. 1). 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. Previous studies found that CAT typically requires no more than half of the items of a given test to obtain equivalent reliability and validity as when the full test is administered in a conventional manner [29,31,32]. Because of this efficiency, CAT is becoming widely used, for example, with the Graduate Record Exam (GRE) [33], with PISA 2018 [34], and recently in physics education research [35].

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 quality of proficiency estimates, especially in terms of accuracy and precision?[1] We represent accuracy and precision by the systematic error and random error. Because the FCI itself is 30 items long, we can choose the length of the FCI-CAT to be anything from 1 to 30 items. In terms of minimizing testing time, the fewer the questions the better. On the other hand, if the test is too short, then the accuracy and precision of analyses is compromised. Our aim is to balance these two factors; namely, we aim to see how short one can make the FCI-CAT with minimal loss of accuracy and precision.

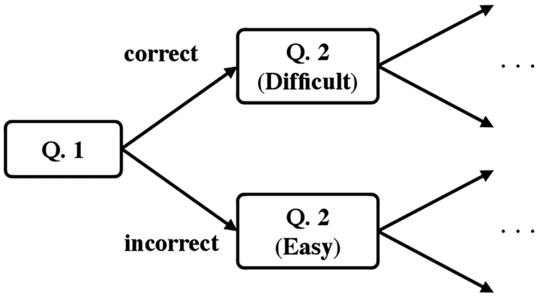


FIG. 1. Example of computerized adaptive testing. In this model of CAT, if a student answers an item correctly (incorrectly), the student will next need to answer a more difficult (easier) item.

To examine the accuracy and precision of the FCI-CAT, like Han *et al.* [26,27], we focus on the pre- and post-group difference measured by the FCI; namely, we compare the accuracy and precision of group difference calculated with the FCI-CAT scores to those calculated with the full FCI scores for a given class size. If the accuracy and precision are comparable, it implies that the FCI-CAT is comparable to the full FCI in regards to measuring pedagogical effects. In our analysis, the measure we focus on is the standardized mean difference (Cohen's d) [37–39]. Cohen's d is defined by the difference of means divided by the pooled standard deviation, and it is used to compare effects across studies in meta-analysis, even when the variables are measured in different ways. DeMars [40] analyzed the accuracy and precision of Cohen's d in the context of item response theory (IRT) with a simulated item bank. We utilize the results of Ref. [40] as a consistency check for our analysis of the FCI-CAT. Although we focus on the accuracy and precision of Cohen's d, there are other criteria (upon which we elaborate in Sec. IV) with which one can examine the optimal length of a CAT-based assessment. As we describe below, by providing a method to analyze optimal length which future researchers of the FCI-CAT can reference, our work serves as a starting point for the FCI-CAT.

The procedure of our analysis is as follows. First, we calibrate (estimate) the item parameters of the 30 FCI items based on the models of IRT [41], which are used in CAT algorithms. Then, to analyze the optimal length of the FCI-CAT, we conduct two simulations that are commonly used in CAT development, a Monte Carlo simulation and a *post hoc* simulation [29,42,43]. Monte Carlo simulations generate responses with pseudorandom numbers, while *post hoc* simulations utilize empirical data (in this case, from the full-length FCI). To ensure that the simulated data are compatible with actual data one might obtain from a real classroom, we examine the consistency of the results of the Monte Carlo and *post hoc* simulations.

The remainder of this paper is organized as follows. In Sec. II, we describe the models of IRT we considered,

[1] Accuracy is the closeness of agreement between a measured value and a true value, and precision is the closeness of agreement between measured values obtained by replicate measurements on similar objects under specified conditions [36].

our CAT settings, our approach to examine optimal test length, and our methods to conduct both the Monte Carlo and *post hoc* simulations. In Sec. III, we present our estimates for item parameters, the validity of the IRT model we used, and the results of the simulations. 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 [44] and RStudio [45]. In addition to the basic package of R, the item parameters of the FCI were calibrated using the package *mirt* [46] and the simulations of the FCI-CAT were conducted using the package *catR* [47,48].

II. METHODOLOGY

A. Models of item response theory

Models of IRT describe the relationship between the latent trait measured by the instrument and the response to an individual item [41]. An advantage of IRT that makes it a popular choice with CAT is that it uses the same scale to place items and respondents, which allows the CAT algorithm to readily match respondents to the item most appropriate for them [29]. Both IRT models and Rasch models (the basic form of which is mathematically identical to the simplest IRT model) have recently been used in the field of physics education research to analyze the structure of assessment tests, especially of the FCI [49–58]. Since a response on the FCI is scored as correct or incorrect (coded as 1 or 0), we consider only the dichotomous models of IRT. Furthermore, we consider only the unidimensional models, for we assume that the latent trait measured by the FCI is dominated by a single proficiency, namely, student conceptual understanding of Newtonian mechanics. This assumption is consistent with the findings of other researchers [49,50]. Furthermore, a unidimensionality test that we describe below confirms that this is an acceptable assumption for our study as well.

Among the dichotomous unidimensional models in IRT, the simplest model is the one-parameter logistic (1PL) model. Based on the 1PL model, the empirical data is fit with the following logistic function:

$$P_i(\theta) = \frac{1}{1+\exp[-(\theta - b_i)]}, \tag{1}$$

where θ represents the proficiency level and $P_i(\theta)$ is the probability that an examinee with a given θ answers the ith question of the FCI correctly. The proficiency distribution in a reference population is often standardized; namely, the estimated mean of θ is set to 0 and the estimated standard deviation of θ is set to 1. In Eq. (1), b_i is the difficulty parameter, which is inversely proportional to the correct answer rate for the item.

The two-parameter logistic (2PL) model is mathematically represented by

$$P_i(\theta) = \frac{1}{1+\exp[-a_i(\theta - b_i)]}, \tag{2}$$

where a_i is the discrimination parameter. The discrimination parameter corresponds to the slope at $\theta = b_i$, where the slope of the curve is steepest. Items with steeper slopes, that is larger a_i, can better distinguish examinees who have different levels of proficiency.

The three-parameter logistic (3PL) model is represented by

$$P_i(\theta) = g_i + \frac{1-g_i}{1+\exp[-a_i(\theta - b_i)]}, \tag{3}$$

where g_i is the asymptotic value of $P_i(\theta)$ when θ approaches negative infinity. It hence represents the probability that an examinee would answer an item correctly by guessing. For example, if respondents with low θ chose randomly among five choices, g_i would approach 0.2.

As we will describe in a later section, our FCI response data cannot be fit sufficiently well to the 1PL model. They can, however, be fit to the 2PL model, as well as to the 3PL model. We examined the optimal length of the FCI-CAT using the 2PL and 3PL models and found better results (that is, the accuracy and precision is better for a given test length and class size) with the 3PL model. Therefore, we describe only our findings with the 3PL model in this paper.

B. Empirical data collection

Our study utilizes two sets of empirical FCI data (Table I). The first empirical dataset (dataset α) is comprised of 2882 responses from Japanese university students. This dataset is used to calibrate the item parameters of the FCI. The second empirical dataset (dataset β) consists of 86 pairs of pretest and post-test responses from students across 3 classes. This dataset is used for our *post hoc* simulation of the FCI-CAT.

Dataset α was collected in Japan from April 2015 to April 2018. The examinees were students at the beginning of introductory physics courses at one public university and

TABLE I. Empirical datasets and their usage.

	N (valid)	Administration	Usage
Empirical dataset α	2882 (2812)	Pre	To calibrate IRT item parameters of the FCI
Empirical dataset β	86 (85) pairs	Pre and Post	To conduct *post hoc* simulations

four private universities. All five of these schools are middle-rank universities in Japan. The total number of survey responses was 2882. From this dataset, we removed the responses of students who did not answer some of the questions, who wrote a letter which was not one of the choices available for a given question, or who wrote the same or serial letters continuously. In total, the number of valid responses was 2812.

Dataset β was collected in Japan from April 2015 to April 2018. The examinees were students of introductory physics courses at a public middle-rank university in Japan. The instructor of the courses utilized interactive-engagement teaching methods, for example, Open Source Tutorials [59]. The FCI was administered at the beginning and the end of the courses. In total, there were 86 pairs of pre- and post-test responses from students across 3 classes. From this, we excluded invalid responses in the same manner as for dataset α. In total, the number of valid pairs was 85.

Most of the examinees in both empirical datasets α and β were first-year students. Most students were in the department of science or the department of technology. The examinees were not given any incentive to participate (in the form of money or extra credit). However, the survey was administered during class, so as to help ensure that students would concentrate on the survey.

C. Settings for computerized adaptive testing

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 [30,47]: (i) initial step, (ii) test step, (iii) stopping step, and (iv) final step. Our settings for the four steps are as follows.

(i) Initial step: In this step, the first item administered to an examinee is selected. As is commonly done, we used the maximum Fisher information (MFI) criterion [30]. The MFI criterion calls for selecting the most informative item for the examinee based upon the current estimate of the proficiency. (Generally speaking, the "most informative item" is the one that will minimize the standard error of the proficiency estimate [41].) When nothing is known about the respondent (as is often the case when the first item is chosen), the information of the items is calculated using the mean proficiency value of the prior population. As is commonly done, we set the prior population mean proficiency value to be zero to have the scale be centered on examinees [29,30].

(ii) Test step: In this step, the proficiency of the examinee is estimated using the current set of item responses and the next item is selected to be administered. As is commonly done, we chose the expected *a posteriori* (EAP) method to estimate the proficiency and the MFI criterion to select the next item [30,47]. The EAP is a common method to estimate the proficiency, and, compared to several alternatives, it has been found to be less biased [40,60] and to be more effective at reducing the test length [61].

(iii) Stopping step: This is the step where the test checks that a certain criterion has been met and the test ends. This criterion is set prior to the test. 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.

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

D. Approach to analyzing optimal test length

As mentioned above, to analyze the optimal length of the FCI-CAT, we varied the test length l from 1 to 30 items and then compared the accuracy and precision of Cohen's d calculated for that l with that calculated for the full FCI ($l = 30$). The population parameter of Cohen's d is given by [37,38]

$$d = \frac{\mu_{\text{post}} - \mu_{\text{pre}}}{\sigma}, \tag{4}$$

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 post-population (we assume that the two population standard deviations are the same, as is done in most parametric data analysis techniques [38]). The numerator of d has the same structure as Rasch gain, which Planinic *et al.* [50,55] as well as Nitta and Aiba [62] suggested to use. The advantage of d is that it compares the group (pre- and post-) proficiencies in relation to σ.

We express the estimator of d on the test length l as $\hat{d}_l$. From within the family of estimators for d, we use the following definition for repeated measures [38,39,63]:

$$\hat{d}_l = \frac{\bar{\theta}^l_{\text{post}} - \bar{\theta}^l_{\text{pre}}}{s_l}, \tag{5}$$

where $\bar{\theta}^l_{\text{pre}}$ and $\bar{\theta}^l_{\text{post}}$ are the means of the final estimated proficiencies of the l-length pre- and post- test, respectively. s_l is the pooled standard deviation for dependent (paired) data defined as

$$s_l^2 = \frac{(s^l_{\text{pre}})^2 + (s^l_{\text{post}})^2 - 2r_l s^l_{\text{pre}} s^l_{\text{post}}}{2(1 - r_l)}, \tag{6}$$

where s^l_{pre} and s^l_{post} are the standard deviations of the final estimated proficiencies of the l-length pre- and post- test, respectively, and r_l is the Pearson correlation coefficient.

Since we simulate the case of class size larger than 40, we do not consider Hedge's correction to the estimator [64].

In our Monte Carlo study (described below), we generate pre- and post-responses to the FCI-CAT and, keeping class size and population parameters fixed, calculate $\hat{d}_l$ 10 000 times for each l to analyze the sampling distribution of $\hat{d}_l$. Following this, we analyze the quality of $\hat{d}_l$, especially in regards to the accuracy and precision. We represent the accuracy and precision by the systematic error and random error, which we measure in terms of the bias and standard error. The measurement error of $\hat{d}_l$ is defined by the difference between $\hat{d}_l$ and the true value of d as [65]

$$e_l = \hat{d}_l - d. \tag{7}$$

The bias of $\hat{d}_l$ is defined by the expected value of e_l as

$$B(\hat{d}_l) = E(e_l) = E(\hat{d}_l) - d, \tag{8}$$

where $E(\hat{d}_l)$ is the expected value of $\hat{d}_l$, estimated by taking the average of 10 000 samples of $\hat{d}_l$. The standard error of $\hat{d}_l$ is given by

$$\mathrm{SE}(\hat{d}_l) = \sqrt{E\{[\hat{d}_l - E(\hat{d}_l)]^2\}}. \tag{9}$$

These quality measures of the estimator $\hat{d}_l$ are summarized by the root-mean-square error (RMSE) defined by the following equation, which equals the square root of the sum of the squared bias and squared standard error:

$$\mathrm{RMSE}(\hat{d}_l) = \sqrt{E[(\hat{d}_l - d)^2]} = \sqrt{B^2 + \mathrm{SE}^2}. \tag{10}$$

With these statistics, we can represent the magnitude of the systematic error and random error, thereby the accuracy and precision. Specifically, we compare the systematic error and random error of the FCI-CAT to those of the full FCI based on the difference of the RMSE, $\mathrm{RMSE}(\hat{d}_l) - \mathrm{RMSE}(\hat{d}_{30})$. If this difference is sufficiently small (see below), we regard the quality of the l-length FCI-CAT to be comparable to that of the full FCI. We are unaware of any other studies that compare bias, SE, and RMSE of Cohen's d obtained from CAT. As such, there are no typical benchmark (or cutoff) values proposed by prior related studies for us to utilize. Researchers who follow this line of work in the future can benefit by comparing their statistics with ours.

E. Procedure of Monte Carlo simulation

An overview of the steps involved in our Monte Carlo study is as follows (see also Table II).

(i) Generate paired pre- and post-true proficiency levels for the simulated respondent (simulee) in a given class size (say, 100 students) with designated population parameters for the true proficiencies (means, common standard deviation, and correlation).

(ii) Generate paired pre- and post-responses to the FCI-CAT for each simulee based on the true proficiencies generated in step 1.

(iii) Using the paired responses, estimate the proficiencies of the simulees and calculate Cohen's d for the class on the test of length l.

(iv) Repeat the above three steps 10 000 times with different random seeds but with fixed class size and population parameters.

(v) Repeat the above four steps with different class sizes.

(vi) Analyze the dependence of the measurement error of $\hat{d}_l$ on the test length and class size to then examine the optimal test length of the FCI-CAT.

Actually, the procedure in our R script deviates somewhat from the above outline for computational efficiency. For example, we actually generated paired pre- and

TABLE II. The first three steps of the Monte Carlo simulation described in Sec. II E. The designated population parameters are mean (μ), standard deviation (σ), and correlation (ρ). θ^j is the true proficiency of the jth simulee. The ith response for jth simulee in the pre- and post- test (u_i^j and v_i^j) take values of 0 (incorrect) or 1(correct). $\hat{\theta}_l^j$ is the estimate of θ^j obtained with the l-length FCI-CAT. $\bar{\theta}^l$ and s_l are the mean and the pooled standard deviation of the estimated proficiencies of the l-length test for class size n.

Step	Input	Output	Process
1.	$\mu_{\text{pre}}, \mu_{\text{post}}, \sigma, \rho$	$(\theta^1_{\text{pre}}, \ldots, \theta^j_{\text{pre}}, \ldots, \theta^n_{\text{pre}})$ $(\theta^1_{\text{post}}, \ldots, \theta^j_{\text{post}}, \ldots, \theta^n_{\text{post}})$	Generate paired pre- and post-true proficiencies for class size n.
2.	θ^j_{pre} θ^j_{post}	$(u^j_1, \ldots, u^j_i, \ldots, u^j_{30})$ $(v^j_1, \ldots, v^j_i, \ldots, v^j_{30})$	Generate pre- and post-responses for the FCI-CAT.
3.	$(u^j_1, \ldots, u^j_l)$ $(v^j_1, \ldots, v^j_l)$	$\hat{\theta}^j_{l,\text{pre}}$ $\hat{\theta}^j_{l,\text{post}}$	Estimate pre- and post-proficiencies of each simulee for the l-length FCI-CAT.
	$(\hat{\theta}^1_{l,\text{pre}}, \ldots, \hat{\theta}^n_{l,\text{pre}})$ $(\hat{\theta}^1_{l,\text{post}}, \ldots, \hat{\theta}^n_{l,\text{post}})$	$\hat{d}_l = \frac{\bar{\theta}^l_{\text{post}} - \bar{\theta}^l_{\text{pre}}}{s_l}$	Calculate the value of Cohen's d.

post-true proficiencies and the subsequent FCI-CAT responses for 100 000 simulees and then resampled 10 000 times for each class size. We explain the details of our procedure in the following paragraphs (see also Appendix for mathematical notes).

We followed a two-step process to generate paired pre- and post-responses. In the first step, we used the function `mvrnorm` of the *MASS* package [66] to generate a pair of proficiencies for a given simulee, one corresponding to the pretest and one corresponding to the post-test. Our simulations consider these values to be the "true" values of the simulee's proficiency, and these are used in predicting how the simulee will answer a given item (see second step, below). We made three assumptions for the population distributions of the pre- and post-true proficiencies. First, we assumed that true proficiency follows a normal distribution, since we found that the distribution of the estimated proficiency for our empirical dataset α (estimated using all 30 FCI items) is unimodal and almost symmetric. Second, we assumed that the standard deviations are the same for both the pre- and post-population distributions, since the standard deviations of the estimated proficiency for the pre- and post-responses of our empirical dataset β took similar values. These assumptions are consistent with Ref. [67] and the articles cited within. A third and final assumption we made was that the pre- and post-true proficiencies are correlated with each other, since the estimated proficiencies for the pre- and post-responses of our empirical dataset β are highly correlated.

We designated typical values for the population parameters (the pre- and post-means, common standard deviation, and correlation) of the bivariate normal population distributions for pre- and post-true proficiencies. Specifically, we chose the parameters such that the estimates by the simulation for the 30-item length test are as close to the statistics calculated with our empirical dataset β as possible (the details are described in Sec. III). These parameters were pretest true proficiency mean $= 0.45$, post-test true proficiency mean $= 0.75$, standard deviation for both sets of true proficiency $= 0.8$, and correlation $= 0.99$. 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 using the function `simulateRespondents` of *catR* package. As discussed above, the CAT algorithm we used selects the next item based upon the MFI criterion. In the simulation, the response to that item is generated based upon the calibrated item parameters of that item and the value of true proficiency for the simulee. For example, suppose that the probability of a correct response for an item is calculated to be 0.75 for a simulee with a given true proficiency [see Eqs. (1)–(3)]. A random number is generated from a uniform distribution within a range of 0 to 1. If the value is 0.75 or less, the generated response is coded as "correct" for that item. If the value is greater than 0.75, then the generated response is "incorrect." The EAP method is used to estimate the proficiency for the respondent, the MFI criterion is then used to choose the next item based upon that estimated proficiency, and the process repeats until all 30 FCI items have simulated responses. To summarize, provided calibrated item parameters and a true proficiency value for a given simulee, the `simulateRespondents` function generates an entire set of correct or incorrect responses and calculates estimated proficiencies for each length (from 1 to 30) of the FCI-CAT. This is done for the simulee both on the pretest and on the post-test (with a different value of true proficiency calculated in the first step, above). In this manner, we generated paired pre- and post-responses and estimated proficiencies for 100 000 simulees for each length of the FCI-CAT.

From the 100,000 paired pre- and post-responses, 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 $\hat{d}_l$ and the corresponding measurement error.

F. Procedure of *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") and examined the consistency of the results with the Monte Carlo study. *Post hoc* simulations are commonly used to determine how short a CAT-based assessment can be without excessively sacrificing accuracy and precision [29,48]. 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 empirical dataset β and the function `simulateRespondents` of the *catR* package. The estimator for the standardized mean difference on the test length l in the *post hoc* simulation is represented as $\hat{d}_l^{\mathrm{ph}}$ and is calculated by Eq. (5), as was done in the Monte Carlo simulation. The estimate for the variance of $\hat{d}_l^{\mathrm{ph}}$ is computed using the formula for matched groups [38],

$$\left(\frac{1}{n}+\frac{(\hat{d}_l^{\mathrm{ph}})^2}{2n}\right)2(1-r_l), \tag{11}$$

where n is the number of pairs.

III. RESULTS

A. Calibration of item parameters

Table III shows the result of the calibration for the item parameters of the FCI based on the 3PL model. In the calibration, we used the empirical dataset α and the function `mirt` of the package *mirt*. From the dataset α, we removed aberrant responses using the standardized person fit index Z_h [68,69], where large negative Z_h values indicate misfit and large positive Z_h values indicate overfit [70]. As is commonly done for the standardized person fit index in Rasch analysis [71], we chose to keep responses with $-2 < Z_h < 2$. In total, we kept 2712 responses. Using these filtered responses, we obtained the estimates and the standard errors in Table III. Since we used a large number of responses, the standard errors of the estimates are quite small.

The estimate of the discrimination parameter (that is, the a parameter) of item 29 is exceptionally small in comparison to the other items. Although an administrator of the FCI-CAT might thus consider eliminating the item, we chose to keep it in our study since we want to examine the optimal length of the FCI-CAT in terms of its accuracy and precision in comparison to that of the full FCI. Since the information function is proportional to the square of the discrimination parameter, the information function of item 29 is relatively small. As we use the MFI criterion in selecting the next item administered, item 29 ends up being one of the last items of the assessment (usually, it is the last item administered). Hence, we expect that its inclusion has no significant effect on our results.

TABLE III. The results of the estimation for the item parameters of the FCI based on the 3PL model ($N = 2712$). The estimates and the standard errors (SEs) for the item parameters are shown.

	$\hat{a}$	SE($\hat{a}$)	$\hat{b}$	SE($\hat{b}$)	$\hat{g}$	SE($\hat{g}$)
Item 1	1.23	0.16	−0.66	0.28	0.14	0.12
Item 2	1.32	0.18	0.35	0.14	0.27	0.05
Item 3	1.16	0.20	−0.88	0.46	0.42	0.14
Item 4	1.10	0.07	0.59	0.05	0.00	0.01
Item 5	1.75	0.15	0.37	0.06	0.06	0.03
Item 6	1.54	0.22	−0.05	0.17	0.46	0.05
Item 7	0.96	0.13	−0.84	0.39	0.02	0.16
Item 8	1.71	0.18	−0.36	0.14	0.34	0.05
Item 9	1.72	0.19	0.12	0.10	0.24	0.04
Item 10	1.69	0.19	−1.13	0.23	0.23	0.13
Item 11	1.38	0.07	0.25	0.04	0.00	0.01
Item 12	1.87	0.25	−0.48	0.17	0.55	0.05
Item 13	3.67	0.28	0.18	0.03	0.04	0.01
Item 14	1.31	0.14	0.27	0.11	0.07	0.04
Item 15	0.61	0.05	0.71	0.09	0.00	0.01
Item 16	1.24	0.07	−0.69	0.05	0.00	0.01
Item 17	1.56	0.14	0.56	0.06	0.01	0.02
Item 18	1.89	0.15	0.11	0.06	0.08	0.03
Item 19	1.25	0.15	−0.20	0.19	0.14	0.08
Item 20	1.26	0.07	−0.31	0.05	0.00	0.01
Item 21	2.13	0.29	0.86	0.06	0.20	0.02
Item 22	2.72	0.32	0.52	0.05	0.28	0.02
Item 23	1.86	0.18	0.47	0.06	0.10	0.03
Item 24	1.58	0.18	−0.31	0.15	0.21	0.07
Item 25	2.76	0.22	0.35	0.04	0.09	0.02
Item 26	2.38	0.17	0.61	0.04	0.03	0.01
Item 27	1.47	0.20	0.27	0.13	0.31	0.05
Item 28	1.20	0.07	−0.60	0.06	0.00	0.02
Item 29	0.20	0.05	−6.00	1.85	0.01	0.14
Item 30	2.88	0.20	0.53	0.03	0.05	0.01

B. Examination of the assumptions of the IRT model

Users of IRT should check that certain assumptions made by the models are satisfied. We confirmed that the assumptions of unidimensionality and local independence at the whole test level are satisfied, and we confirmed that the IRT model we use is a good fit for our data.

1. Unidimensionality and local independence

We examined the unidimensionality of the FCI via a principal component analysis with the tetrachoric correlation matrix [41] using the filtered empirical dataset α. Figure 2 shows the scree plot for the eigenvalues of the correlation matrix. The vertical axis measures the eigenvalue of the components labeled on the x axis. The first eigenvalue (far left on the x axis) is about 5 times larger than the rest, which suggests there is one dominant dimension and it is reasonable to assume unidimensionality for our FCI dataset as was done in Ref. [49].

Since our dataset is unidimensional, it can be argued that the assumption of overall local independence must be satisfied as well [49]. In addition, we also explicitly evaluated the local independence assumption by using Yen's Q_3 statistic [72]. Yen's Q_3 is defined as the correlation of the residuals between each pair of items.

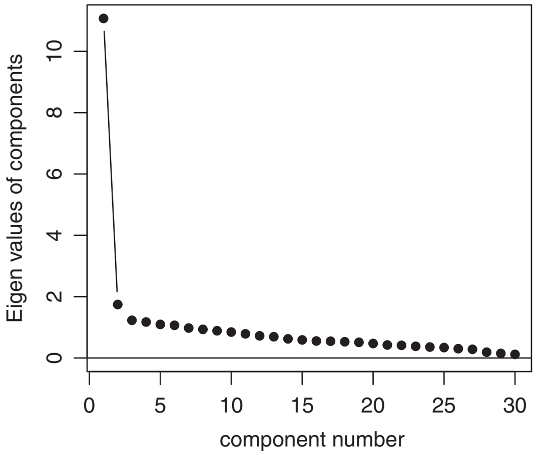


FIG. 2. The scree plot for the tetrachoric correlation matrix via a principal component analysis using the filtered empirical dataset α ($N = 2712$).

It is common to flag the pair of items as being locally dependent if the absolute value of Q_3 is larger than 0.2 [73]. In the present study, we found that the values of Q_3 for 4 pairs (out of a total of 435 item pairs) were flagged in the 3PL model. These values and the item pairs are 0.203 (1–2), 0.224 (5–18), 0.230 (8–9), 0.310 (23–24). Although these item pairs may violate local independence, since there were only a small number of such pairs (4/435), and since the magnitudes of the Q_3 values are not so large, our assumption of local independence is by and large supported [74]. Furthermore, it can be argued that, as was done in Ref. [75], since the mean value of Q_3 in the present study is small (e.g., –0.027 for the 3PL model), our dataset has sufficient local independence at the whole test level. However, in future studies, we recommend separating these items either by dropping one item from each dependent pair or combining each high-Q_3 pair into a single polytomous item as was done in Ref. [76]. Doing so will help both ensure that items are locally independent and to improve estimations.

2. Goodness of fit and model selection

We evaluated the goodness of fit of 1PL, 2PL, and 3PL IRT models to the response data with the standardized root mean square residual (SRMSR) [77]. We calculated the values of SRMSR for the models using the filtered empirical dataset α. A value of SRMSR less than 0.05 is considered to indicate that the model is well fitted [77]. We found that the SRMSR is 0.079 for the 1PL model, 0.041 for the 2PL model, and 0.041 for the 3PL model. This result indicates that the 2PL and 3PL models fit the data well; the 1PL model, on the other hand, does not.

In choosing a model, we want to fit the response data as closely as possible while using minimal parameters. The balance of these conditions can be examined using the Bayesian information criterion (BIC) [78]. BIC increases if the deviance of the model from the data increases and if the number of parameters increases; thus, the model with the lowest BIC is the most preferable. Using the filtered empirical dataset α, we found that BIC is 9.22×10^4 for the 1PL model, 9.00×10^4 for the 2PL model, and 9.00×10^4 for the 3PL model. This finding indicates that BICs of the 2PL and 3PL model are comparable, and smaller than that of the 1PL model. As we mentioned above, since we obtained greater accuracy and precision with the 3PL model, we describe only our findings with the 3PL model in the following analysis.

C. Results of the Monte Carlo study

1. Descriptive results

All results in this and the following section (Sec. III D) are based on the 3PL model. As we described in Sec. II E, we first chose population parameters (left column of Table IV) such that the estimates produced by the simulations when all 30 FCI questions are used (middle column) are as close to the statistics of our empirical data (right column) as possible.

TABLE IV. We chose population parameters of true proficiency for the Monte Carlo simulation (left column) such that the parameters estimated by the simulation for the full-length FCI (center column, $N = 100\,000$) would be similar to the parameters measured by the empirical data β (right column, $N = 78$).

	Population parameters		Monte Carlo estimates		Empirical estimates	
	Pre	Post	Pre	Post	Pre	Post
Mean	0.45	0.75	0.41	0.68	0.40	0.68
Standard deviation	0.80	0.80	0.81	0.78	0.82	0.79
Correlation	0.99		0.85		0.85	
Cohen’s *d*	0.38		0.34		0.35	

Figure 3 shows a typical example trend (one of the 10 000 samplings) of $\hat{d}_l$ when the class size equals 80 and the true value of Cohen’s d is 0.38. As shown, $\hat{d}_l$ first approaches $\hat{d}_{30}$ at $l \sim 10$, although there are fluctuations thereafter. Again, Fig. 3 is just one example out of the 10 000 simulated cases. We illustrate how $\hat{d}_l$ varies from case to case in Fig. 4, which shows the sampling distribution of $\hat{d}_l$ at $l = 10$ for the same class size and true d as in Fig. 3. From the histogram, we can see that $\hat{d}_{10}$ is close to being a normal distribution (a result we found for any test length). The bias, standard error, and RMSE of the distribution in Fig. 4 was calculated as $B(\hat{d}_{10}) = -0.043$, $\mathrm{SE}(\hat{d}_{10}) = 0.076$, and $\mathrm{RMSE}(\hat{d}_{10}) = 0.088$. In the following section, we analyze how these measures depend on the test length and the class size. Then we discuss how we should interpret the magnitude of these values.

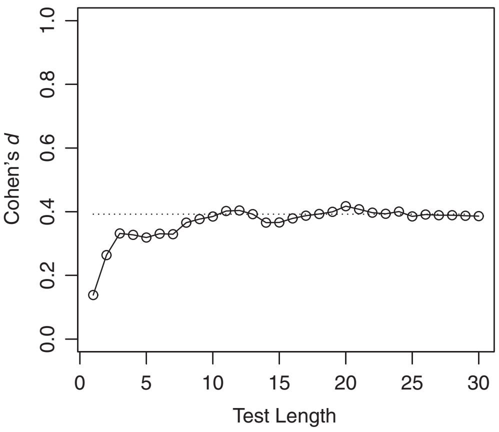


FIG. 3. The typical trend of Cohen’s d as a function of test length of the FCI-CAT, determined by Monte Carlo simulation (class size = 80). The true value of Cohen’s d is 0.38 as shown by a dotted line.

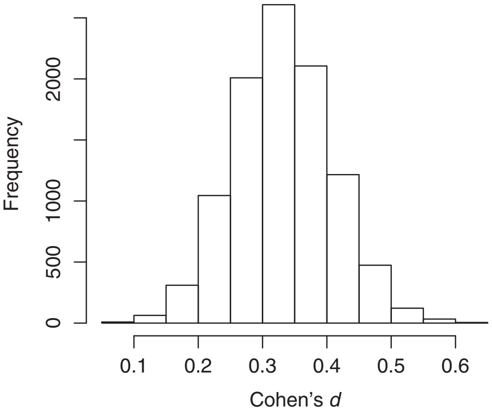


FIG. 4. The simulated sampling distribution of Cohen's d for a test length of 10 items and class size of 80 (sample size $= 10\,000$ classes). The true value of Cohen's d is 0.38.

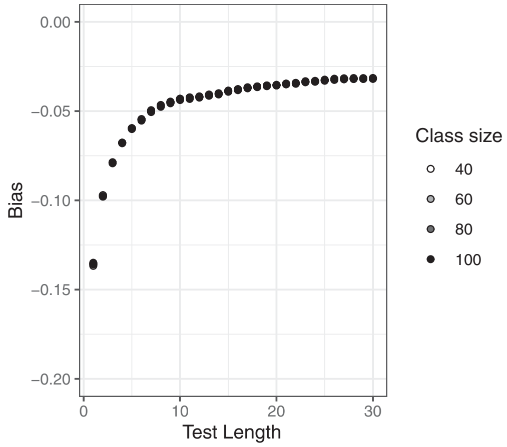
2. Test length and class size dependence of the statistics

Figure 5 shows the test length dependence of the bias $B(\hat{d}_l)$, which was calculated by Eq. (8) with 10 000 samples of $\hat{d}_l$ for the true d of 0.38. The figure also shows the class size dependence of $B(\hat{d}_l)$ for the class size of 40, 60, 80, and 100. Three major observations on the bias can be made from the figure. First, the dependence on the class size is too small to be visible in the figure (the differences in bias are less than 0.001 for any given test length). Second, the absolute value of the bias decreases as the test length increases but it is somewhat less than zero even if $l = 30$. In the case when $n = 40$, $B(\hat{d}_{30})$ is -0.031 at $l = 30$. This result is consistent with the study by DeMars [40], which showed that the bias of Cohen's d of a 30-item test reaches -0.02 when $d = 0.2$, and -0.04 when $d = 0.5$ using the 3PL model and EAP method. Third, $B(\hat{d}_l)$ gets close to $B(\hat{d}_{30})$ as the test length l increases. In the case when $n = 40$, the difference $B(\hat{d}_l) - B(\hat{d}_{30})$ is -0.028 at $l = 5$, -0.012 at $l = 10$, and -0.007 at $l = 15$.

Figure 6 shows test length dependence of the standard error $\mathrm{SE}(\hat{d}_l)$, which was calculated by Eq. (9) with 10 000 samples of $\hat{d}_l$ for the true d of 0.38. Three major observation on the $\mathrm{SE}(\hat{d}_l)$ can be made from the figure. First, the $\mathrm{SE}(\hat{d}_l)$ decreases as the test length increases. Second, the $\mathrm{SE}(\hat{d}_l)$ decreases as the class size increases. The dependence for the class size n is $1/\sqrt{n}$, as expected from Eq. (11). Third, $\mathrm{SE}(\hat{d}_l)$ gets close to $\mathrm{SE}(\hat{d}_{30})$ as the test length increases. For example, when the class size is 40, the difference $\mathrm{SE}(\hat{d}_l) - \mathrm{SE}(\hat{d}_{30})$ is 0.030 at $l = 5$, 0.014 at $l = 10$, and 0.007 at $l = 15$.

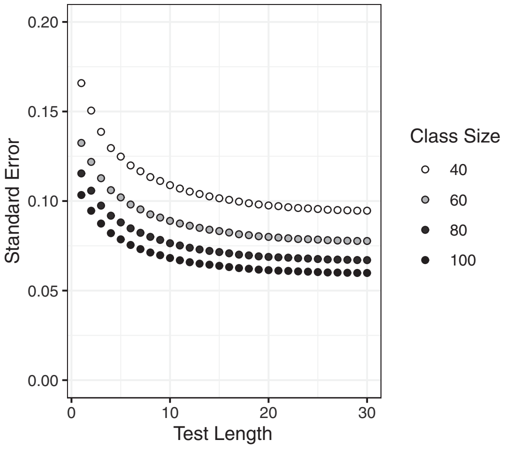
3. Examining the optimal test length of the FCI-CAT

Finally, we examine the optimal test length of the FCI-CAT based on the root-mean-square error of Cohen's d.



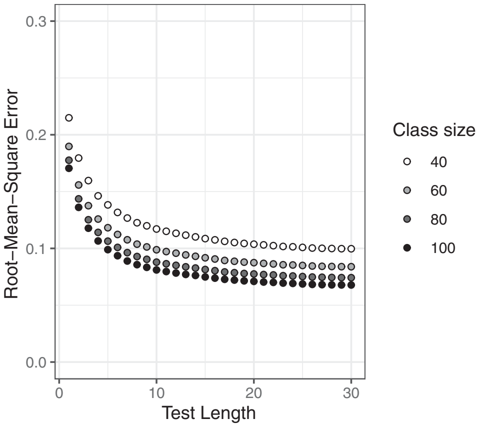
Class size	Test length (l)					
(n)	5	10	15	20	25	30
40	−0.060	−0.043	−0.039	−0.035	−0.032	−0.031
60	−0.060	−0.043	−0.039	−0.035	−0.033	−0.032
80	−0.060	−0.043	−0.039	−0.036	−0.033	−0.032
100	−0.060	−0.044	−0.039	−0.036	−0.033	−0.032

FIG. 5. The bias of Cohen's d as a function of test length of the FCI-CAT, as calculated by Monte Carlo simulation [class size $= (40, 60, 80, 100)$, sample size $= 10\,000$ classes]. The true value of Cohen's d is 0.38. The values in the table are characteristic points plotted in the graph.



Class size	Test length (l)					
(n)	5	10	15	20	25	30
40	0.125	0.109	0.102	0.098	0.096	0.095
60	0.102	0.089	0.083	0.080	0.078	0.078
80	0.088	0.076	0.072	0.069	0.068	0.067
100	0.079	0.068	0.064	0.061	0.060	0.060

FIG. 6. The standard error of Cohen's d as a function of test length of the FCI-CAT, as calculated by Monte Carlo simulation [class size $= (40, 60, 80, 100)$, sample size $= 10\,000$ classes]. The true value of Cohen's d is 0.38. The values in the table are characteristic points plotted in the graph.



Class size		Test length (l)					
(n)		5	10	15	20	25	30
40	RMSE	0.138	0.117	0.109	0.104	0.101	0.100
	%Inc	38.7	17.4	9.0	4.0	1.2	–
60	RMSE	0.118	0.099	0.092	0.087	0.085	0.084
	%Inc	40.9	17.8	9.4	4.2	1.3	–
80	RMSE	0.106	0.088	0.081	0.078	0.075	0.074
	%Inc	43.3	18.4	9.7	4.4	1.3	–
100	RMSE	0.099	0.081	0.075	0.071	0.069	0.068
	%Inc	45.9	19.5	10.5	4.7	1.4	–

FIG. 7. The root-mean-square error of Cohen's d as a function of test length of the FCI-CAT, as calculated by Monte Carlo simulation [class size $= (40, 60, 80, 100)$, sample size $= 10\,000$ classes]. The true value of Cohen's d is 0.38. The values in the table are characteristic points plotted in the graph. The percent increase of the RMSE from the full-length test is denoted by %Inc.

Figure 7 shows test length dependence of the RMSE, which was calculated by Eq. (10) with 10 000 samples of $\hat{d}_l$ for the true d of 0.38. From the figure, we can see that the trend of $\mathrm{RMSE}(\hat{d}_l)$ is similar to that of $\mathrm{SE}(\hat{d}_l)$, namely, it decreases as the test length increases and as the class size increases.

Since there is no typical benchmark for the magnitude of RMSE, assuming that the RMSE of the full-length test is sufficient, we analyze the percent increase of the RMSE from the full-length test, $\%\mathrm{Inc} = [\mathrm{RMSE}(\hat{d}_l) - \mathrm{RMSE}(\hat{d}_{30})]/\mathrm{RMSE}(\hat{d}_{30}) \times 100$. The result is summarized in the table of Fig. 7.

Remember that our objective is to examine the test length of the FCI-CAT in which the accuracy and precision of the l-length test are comparable to those of the full-length test for a given class size. As an illustration, for the class sizes $n = 40$ and $n = 100$, we examine the optimal test length of the FCI-CAT as follows. If we take 10% as a cutoff for the percent increase of the RMSE, the optimal test length is 15 items ($\mathrm{RMSE} = 0.109$, $\%\mathrm{Inc} = 9.0$) for $n = 40$ and 16 items ($\mathrm{RMSE} = 0.074$, $\%\mathrm{Inc} = 8.9$) for $n = 100$. Similarly, if we take 5% as a cutoff, the optimal test length is 19 items ($\mathrm{RMSE} = 0.104$, $\%\mathrm{Inc} = 4.7$) for $n = 40$ and 20 items ($\mathrm{RMSE} = 0.071$, $\%\mathrm{Inc} = 4.7$) for $n = 100$. The decision of whether to use 10% or 5% (or some other value) as the cutoff of the percent increase may be determined by the researcher in coordinating with the teacher regarding available testing time.

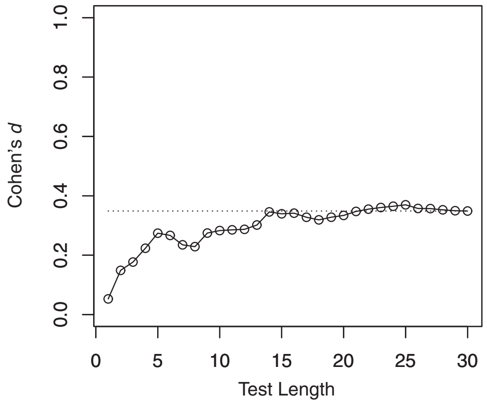
Note that the case of $n = 100$ requires slightly more test items, although the RMSE of $n = 100$ is smaller than that of $n = 40$ at the same test length. This is because the RMSE of the full-length test (denominator of the percent increase) is larger for the smaller class size. This could be problematic, depending upon a researcher's goals, but since our objective is to examine the test length with which the accuracy and precision of the l-length test are comparable to that of the full-length test *for each class size*, this has no effect on the merit of our findings. This issue could be resolved were we to use a fixed denominator of the percent increase instead of $\mathrm{RMSE}(\hat{d}_{30})$ as our measure of error. However, defining such a denominator is beyond the scope of this paper. An additional alternative for a denominator would be the true value of Cohen's d itself as in Ref. [65]. In our case, the true value of d was equal to 0.38 and so this would be potentially useful as a measure of error. However, this could not be used as a general approach, as it is possible for d to be very close to zero (when post-test scores are the same as pretest scores), and the ratio would diverge.

We can summarize the above results as follows. For a class size of 40, we may reduce the test length of the FCI-CAT to 15–19 items, thereby reducing the test time of the FCI to 50%–63%, with an accompanying decrease in accuracy and precision of only 5%–10%. For a larger class size of 100, the result is very similar (16–20 items are needed). It is important to note that these findings are specific to a given value of true Cohen's d, which was determined from our empirical data. Additional research is necessary to see how the results are different for other student populations.

D. Results of *post hoc* study

As described above, we conducted an additional simulation using empirical responses (a *post hoc* simulation) to compare with the results from the Monte Carlo study. First, from the dataset β, we removed aberrant responses using the standardized person fit index Z_h [68] as was done for dataset α, resulting in 78 pairs of preresponses and postresponses.

Figure 8 shows Cohen's d calculated with this data as a function of the length of the FCI-CAT. Upon inspection, this graph appears similar to Fig. 3, and this similarity is expounded by numerical values in the table of Fig. 8. As there are 78 students in dataset β, we compare simulation results with what was obtained with the Monte Carlo simulation for a class size of 80. The absolute difference between the estimate $\hat{d}_l^{\mathrm{ph}}$ and the



Test length	5	10	15	20	25	30
$\hat{d}_{30}^{\mathrm{ph}}$	0.274	0.283	0.339	0.334	0.370	0.349
$\mathrm{E}(\hat{d}_l)(n=80)$	0.315	0.332	0.336	0.339	0.342	0.343
$\left\lvert \hat{d}_{30}^{\mathrm{ph}} - \mathrm{E}(\hat{d}_l) \right\rvert$	0.041	0.049	0.003	0.005	0.028	0.006
$\mathrm{SE}(\hat{d}_l^{\mathrm{ph}})$	0.074	0.070	0.067	0.066	0.066	0.063
$\mathrm{SE}(\hat{d}_l)(n=80)$	0.088	0.076	0.072	0.069	0.068	0.067

FIG. 8. Cohen's d as a function of test length of the FCI-CAT, as calculated by *post hoc* simulation (sample size $= 78$ respondents). The dotted line shows $d = 0.35$, which is the value of Cohen's d at a test length of 30. The values in the table are characteristic points plotted in the graph and associated values calculated from the Monte Carlo simulation (see description below).

expected value $E(\hat{d}_l)$ calculated with the Monte Carlo simulation for $d = 0.38$, $n = 80$ is less than the standard error $\mathrm{SE}(\hat{d}_l^{\mathrm{ph}})$ calculated by Eq. (11). In addition, the estimated standard error $\mathrm{SE}(\hat{d}_l^{\mathrm{ph}})$ takes similar values to the $\mathrm{SE}(\hat{d}_l)$ calculated with the Monte Carlo simulation for $d = 0.38$, $n = 80$.

From these results, we can see that the results of the Monte Carlo study and the *post hoc* simulation are consistent. This consistency supports the adequacy of our Monte Carlo study and its relevance in terms of conducting the FCI-CAT in real classrooms.

IV. DISCUSSION

A. Summary

In order to shorten the test time of the FCI, we examined the optimal length of a computerized adaptive testing (CAT)-based version of the FCI. Using a Monte Carlo simulation, we analyzed the bias, standard error, and root-mean-square error of Cohen's d of the l-length FCI-CAT to find values that are comparable to those obtained when all 30 items of the FCI are used. We also conducted a *post hoc* simulation to examine the consistency of the results from this simulated data with what would have been obtained in an actual classroom. As a result, we found, that for a class size of 40, we may reduce the test length of the FCI-CAT to 15–19 items, thereby reducing the test time of the FCI to 50%–63%, with an accompanying decrease in accuracy and precision of only 5%–10%. The result is almost identical for a larger class size of 100. Note that the threshold of 5%–10% is one that we arbitrarily chose, as there are no prior reference values published in literature. Teachers or education researchers faced with more limited class time may opt to more dramatically reduce the length of the FCI, accepting the greater decrease in accuracy and precision. The results of the Monte Carlo study and the *post hoc* simulation were consistent, which supports the adequacy of our Monte Carlo study and its relevance in terms of conducting the FCI-CAT in real classrooms. It is important to note that these findings are specific to a given value of true Cohen's d, which was determined from our empirical data. Additional research is necessary to see how the results are different for other student populations.

B. Limitations and future work

As we mentioned in the introduction, we have provided a starting point for the FCI-CAT. There are many options to explore that may improve the effectiveness (that is, shorten the test time) of the FCI-CAT. For example, using other models in IRT (e.g., multidimensional models [51,54,57,79]) in the parameter estimations and/or using other algorithms in the CAT steps (e.g., using the precision criterion instead of the length criterion in the stopping step) may allow for a shortening of the test. The FCI-CAT can be improved in other means as well. For example, one may wish to prioritize content balancing [30] to ensure that the same set of concepts is covered in the FCI-CAT as in the original FCI. The FCI can also be improved by removing gender unfair items [53,80–82] or by dropping one item from each locally dependent pair [76]. The use of the FCI-CAT instead of the full length FCI also invites work concerning reducing item exposure [30], to reduce the risk that the FCI items leak into the public sphere, thereby compromising the instrument. In order to prevent item parameter drift [83] and maintain the item pool for longitudinal testing (e.g., pre- and post- testing), it is necessary to examine the CAT algorithm to control for item exposure. These concerns are important aspects of instrument validity, and they must also be considered. Doing so, however, may require a longer test length than what our analysis has suggested. 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.

In our work, we have utilized a Monte Carlo simulation to measure the quality of the FCI-CAT, focusing on the accuracy and precision of the effect size; however, there are other statistics that could alternatively be used. For example, it is possible to consider the coverage of the confidence intervals [84] of the effect size instead of the point estimators as we have done. Furthermore, there are other criteria one can use to compare a CAT-based assessment with a conventional test, such as reliability and validity [31,35].

We are also interested in how the FCI-CAT can be optimized for analyzing the pre- and post-test difference in the performance of individual students. For such a study, since Cohen's d is an effect size defined for the group difference, a different measure (one appropriate for use with single cases) is necessary. Using such a statistic, it would be possible to examine, on a student-by-student basis, the minimal test length without excessively compromising the accuracy and precision of the score. Instead of using the fixed length criterion, another possible approach to evaluate individual change is using the classification criterion for the stopping step of the FCI-CAT [30]. With this criterion, the items are administered until the provisional confidence interval of the current proficiency estimate no longer overlaps a predefined proficiency threshold, so that we can evaluate whether an individual student has a proficiency level greater or less than the proficiency level of interest. For example, we could choose the threshold to be the point at which students are considered to have Newtonian Mastery (this threshold for the paper-and-pencil FCI is the score of 85% [2]). If too few items are administered, such that the confidence interval overlaps this threshold, there would be insufficient certainty to make conclusions about the students' mastery of Newtonian mechanics. Sufficient items should be administered such that the confidence interval does not overlap the classification threshold and the student can be classified as having mastered Newtonian mechanics (or not). These are only some of the possible approaches for analyzing individual change between pre- and post-instruction.

Concurrently with the analyses we reported here, we have conducted a trial administration of the FCI-CAT to Japanese students. 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 greater concentration of students, since instructors can monitor the students during the test. After this trial, we interviewed a number of the examinees to find any problems with the test, for example, in regards to the interface. We next plan to administer the FCI-CAT with real classes of students both pre and post-semester to analyze the effect size distribution. We will compare these results with what we discussed above from our Monte Carlo simulation.

ACKNOWLEDGMENTS

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

APPENDIX: MATHEMATICAL NOTE FOR THE PARAMETRIZATION OF PROFICIENCY

We define the two-dimensional random vector $\mathit{\Theta}$, which contains the random variables to represent the pre- and post-true proficiencies as

$$\mathit{\Theta} = \begin{pmatrix} \Theta_{\mathrm{pre}} \\ \Theta_{\mathrm{post}} \end{pmatrix}.$$

The random vector $\mathbf{\Theta}$ follows a bivariate normal distribution which is represented as

$$\mathbf{\Theta} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

where $\boldsymbol{\mu}$ is a two-dimensional mean vector and $\mathbf{\Sigma}$ is a 2×2 covariance matrix,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{\mathrm{pre}} \\ \mu_{\mathrm{post}} \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma^2_{\mathrm{pre}} & \rho\sigma_{\mathrm{pre}}\sigma_{\mathrm{post}} \\ \rho\sigma_{\mathrm{pre}}\sigma_{\mathrm{post}} & \sigma^2_{\mathrm{post}} \end{pmatrix}.$$

We parametrize the standardized mean difference by

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

where we assumed $\sigma_{\mathrm{pre}} = \sigma_{\mathrm{post}} = \sigma$. The estimator of the standardized mean difference is represented by

$$\hat{d}(\mathbf{\Theta}) = \frac{\bar{\Theta}_{\mathrm{post}} - \bar{\Theta}_{\mathrm{pre}}}{S_{\mathrm{pl}}},$$

where for the class size of n,

$$\begin{aligned}
\Theta_{\mathrm{pre}} &= (\Theta^1_{\mathrm{pre}}, \dots, \Theta^n_{\mathrm{pre}}), \\
\Theta_{\mathrm{post}} &= (\Theta^1_{\mathrm{post}}, \dots, \Theta^n_{\mathrm{post}}), \\
\bar{\Theta}_{\mathrm{pre}} &= \frac{\Theta^1_{\mathrm{pre}} + \cdots + \Theta^n_{\mathrm{pre}}}{n}, \\
\bar{\Theta}_{\mathrm{post}} &= \frac{\Theta^1_{\mathrm{post}} + \cdots + \Theta^n_{\mathrm{post}}}{n}, \\
S^2_{\mathrm{pl}} &= \frac{S^2_{\mathrm{pre}} + S^2_{\mathrm{post}} - 2R_\theta S^2_{\mathrm{pre}} S^2_{\mathrm{post}}}{2(1 - R_\theta)}, \\
S^2_{\mathrm{pre}} &= \frac{1}{n-1} \sum_{k=1}^{n} (\Theta^k_{\mathrm{pre}} - \bar{\Theta}_{\mathrm{pre}})^2, \\
S^2_{\mathrm{post}} &= \frac{1}{n-1} \sum_{k=1}^{n} (\Theta^k_{\mathrm{post}} - \bar{\Theta}_{\mathrm{post}})^2, \\
R_\theta &= \frac{\mathrm{Cov}(\Theta_{\mathrm{pre}}, \Theta_{\mathrm{post}})}{S_{\mathrm{pre}} S_{\mathrm{post}}}.
\end{aligned}$$

We define $\boldsymbol{\theta}$ as the value of the random vector $\boldsymbol{\Theta}$ to be estimated in the observation. Using the l-length FCI-CAT with EAP method, we obtain the estimates $\hat{\boldsymbol{\theta}}_l$ for $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}}_l = \begin{pmatrix} \hat{\boldsymbol{\theta}}^l_{\text{pre}} \\ \hat{\boldsymbol{\theta}}^l_{\text{post}} \end{pmatrix}.$$

Then we calculate the estimates for d with $\hat{d}_l(\boldsymbol{\Theta} = \hat{\boldsymbol{\theta}}_l)$.

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

[2] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[3] E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley & Sons, Hoboken, NJ, 2003).

[4] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).

[5] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).

[6] 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).

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

[8] A. Savinainen and J. Viiri, The Force Concept Inventory as a Measure of Students Conceptual Coherence, Int. J. Sci. Math. Educ. **6**, 719 (2008).

[9] 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).

[10] 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).

[11] 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).

[12] K. F. Wilson and D. J. Low, On second thoughts…: Changes of mind as an indication of competing knowledge structures, Am. J. Phys. **83**, 802 (2015).

[13] D. J. Low and K. F. Wilson, The role of competing knowledge structures in undermining learning: Newton's second and third laws, Am. J. Phys. **85**, 54 (2017).

[14] T. F. Scott and D. Schumayer, Conceptual coherence of non-Newtonian worldviews in Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **13**, 010126 (2017).

[15] 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).

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

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

[18] P. Eaton, K. Vavruska, and S. Willoughby, Exploring the preinstruction and postinstruction non-Newtonian world views as measured by the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 010123 (2019).

[19] D. P. Waters, D. Amarie, R. A. Booth, C. Conover, and E. C. Sayre, Investigating students' seriousness during selected conceptual inventory surveys, Phys. Rev. Phys. Educ. Res. **15**, 020118 (2019).

[20] 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).

[21] 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).

[22] 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).

[23] 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).

[24] 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.

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

[26] 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).

[27] J. Han, K. Koenig, L. Cui, J. Fritchman, D. Li, W. Sun, Z. Fu, and L. Bao, Experimental validation of the half-length Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **12**, 020122 (2016).

[28] A. Madsen and S. McKagan, Administering research-based assessments online, PhysPort, Expert Recommendations, https://www.physport.org/recommendations/Entry.cfm?ID=93329 (2020).

[29] N. Thompson and D. Weiss, A framework for the development of computerized adaptive tests, Pract. Assess. Res. Eval. **16**, 1 (2011), https://scholarworks.umass.edu/pare/vol16/iss1/1/.

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

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

[32] D. J. Weiss and G. G. Kingsbury, Application of computerized adaptive testing to educational problems, J. Educ. Measure. **21**, 361 (1984).

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

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

[35] 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).

[36] International Vocabulary of Metrology Basic and General Concepts and Associated Terms (VIM 3rd edition), Tech. Rep., Joint Committee for Guides in Metrology, 2012.

[37] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, 1988).

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

[39] D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs, Frontiers Psychol. **4**, 863 (2013).

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

[41] C. DeMars, *Item Response Theory* (Oxford University Press, New York, 2010).

[42] M. Harwell, C. A. Stone, T.-C. Hsu, and L. Kirisci, Monte Carlo studies in item response theory, Appl. Psychol. Meas. **20**, 101 (1996).

[43] B. Erdem-Kara, Computer adaptive testing simulations in R, Int. J. Assess. Tools Educ. **6**, 44 (2019), https://dergipark.org.tr/en/pub/ijate/issue/43543/621157.

[44] R. C. Team, R: A Language and Environment for Statistical Computing (2019).

[45] R. Team, RStudio: Integrated Development Environment for R (2019).

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

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

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

[49] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[50] 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).

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

[52] S. Rakkapao, S. Prasitpong, and K. Arayathanitkul, Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique, Phys. Rev. Phys. Educ. Res. **12**, 020135 (2016).

[53] 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).

[54] 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).

[55] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, Rasch analysis in physics education research: Why measurement matters, Phys. Rev. Phys. Educ. Res. **15**, 020111 (2019).

[56] 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).

[57] 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).

[58] P. Eaton and S. Willoughby, Identifying a preinstruction to postinstruction factor model for the Force Concept Inventory within a multitrait item response theory framework, Phys. Rev. Phys. Educ. Res. **16**, 010106 (2020).

[59] R. E. Scherr and A. Elby, Enabling Informed Adaptation of Reformed Instructional Materials, AIP Conf. Proc. **883**, 46 (2007).

[60] H. Wainer and D. Thissen, Estimating ability with the wrong model, J. Educ. Stat. **12**, 339 (1987).

[61] A. K. Okan Bulut, Application of computerized adaptive testing to entrance examination for graduate studies in Turkey, Eurasian J. Educ. Res. **49**, 61 (2012), https://ejer.com.tr/en/archives/2012-fall-issue-49/application-of-computerized-adaptive-testing-to-entrance-examination-for-graduate-studies-in-turkey-2712.

[62] H. Nitta and T. Aiba, An alternative learning gain based on the Rasch model, Phys. Educator **01**, 1950005 (2019).

[63] J. M. Nissen, R. M. Talbot, A. Nasim Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 010115 (2018).

[64] J.-C. Goulet-Pelletier and D. Cousineau, A review of effect sizes and their confidence intervals, Part I: The Cohen's d family, Quant. Methods for Psychol. **14**, 242 (2018).

[65] J. Bendat and A. Piersol, *Random Data: Analysis and Measurement Procedures*, 4th ed. (Wiley, Hoboken, 2010).

[66] W. N. V. Ripley and B. D., *Modern Applied Statistics with S*, 4th ed. (Springer, New York, 2002).

[67] W.-C. Wang and H.-C. Chen, The standardized mean difference within the framework of item response theory, Educ. Psychol. Meas. **64**, 201 (2004).

[68] F. Drasgow, M. V. Levine, and E. A. Williams, Appropriateness measurement with polychotomous item response models and standardized indices, Brit. J. Math. Stat. Psychol. **38**, 67 (1985).

[69] R. R. Meijer, A. S. M. Niessen, and J. N. Tendeiro, A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program, Assessment **23**, 52 (2016).

[70] C. D. Desjardins and O. Bulut, *Handbook of Educational Measurement and Psychometrics Using R*, 1st ed. (CRC Press, Boca Raton, 2018).

[71] T. G. Bond and C. M. Fox, *Applying the Rasch Model*, 3rd ed. (Routledge, New York, 2015).

[72] 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).

[73] W.-H. Chen and D. Thissen, Local dependence indexes for item pairs using item response theory, J. Educ. Behav. Stat. **22**, 265 (1997).

[74] L. Ding, Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis, Phys. Rev. ST Phys. Educ. Res. **10**, 010105 (2014).

[75] L. Bao, Y. Xiao, K. Koenig, and J. Han, Validity evaluation of the Lawson classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020106 (2018).

[76] 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).

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

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

[79] W.-C. Wang and W. Chyi-In, Gain score in item response theory as an effect size measure, Educ. Psychol. Meas. **64**, 758 (2004).

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

[84] T. M. Carsey and J. J. J. Harden, *Monte Carlo Simulation and Resampling Methods for Social Science*, 1st ed. (SAGE, Thousand Oaks, CA, 2013).

[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).