Balancing content of computerized adaptive testing for the Force Concept Inventory

Jun-ichiro Yasuda
Institute of Arts and Sciences, Yamagata University, 1-4-12 Kojirakawa, Yamagata, Yamagata, 990-8560, Japan

Michael M. Hull
Austrian Educational Competence Centre for Physics,
University of Vienna, Porzellangasse 4/2/2, Vienna , Vienna, 1090, Austria

As a method to shorten the test time of the Force Concept Inventory (FCI), we previously suggested 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. As a step to develop a CAT-based version of the FCI (FCI-CAT), we previously examined the optimal test length of the FCI-CAT such that accuracy and precision [which were measured in terms of root-mean-square error (RMSE)] of Cohen's d would be comparable to that of the full FCI for a given class size. The objective of this paper is to address an issue in our previous study to improve the FCI-CAT. We consider content balancing 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. Doing so ensures that items from each subgroup are administered. We conducted a Monte Carlo simulation to analyze how implementing an algorithm of content balancing affects the RMSE of Cohen's d. As a result, we found that, for a class size of 40, the increase of the RMSE due to content balancing is 6%–7% for test lengths of 2–5 items and less than 1% if the test length is larger than 13 items. This result indicates that for a sufficiently large test length (say, larger than 13 items), content balancing does not compromise the accuracy and precision of the FCI-CAT. Hence, we recommend that the FCI-CAT incorporate content balancing, provided the test length is larger than 13 items.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.Yasuda

I. INTRODUCTION

The Force Concept Inventory (FCI) [1] is one of the most widely used research-based assessments in physics education. 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 to 30 min to complete the test. By administering the FCI both before and after instruction, we can measure the effects of that instruction in terms of improvement of students' scores.

When administering an assessment like the FCI in a classroom, the shorter the test time, the better. Many instructors are likely to be reluctant to carve time out of their crowded schedules to administer the assessment [2]. To reduce the test time, Han *et al.* [3] 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 [2, 4].

Recently, we [5] 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 so doing, unnecessarily easy items and unnecessarily hard items can be avoided, resulting in the test length being shortened in comparison to standard test administration [6]. Because of its efficiency, CAT is becoming more and more widely used, for example, with PISA 2018 [7], and it has recently been introduced to physics education research as well [8].

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 [9]. Previously, we [5] focused on the accuracy and precision of the standardized mean difference [10], a statistic to quantify the pre- and post-group difference. 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 objective of this paper is to address an issue in our previous study [5] to improve the FCI-CAT. The issue is considering *content balancing* [6, 11], ensuring that the same set of concepts assessed in the original test is covered in the CAT administration for each respondent. Although the FCI has a unidimensional structure [12], it is not "perfectly" unidimensional, and there is room for content balancing to improve the FCI-CAT. One additional benefit of content balancing is to increase acceptance of adaptive testing by practitioners and decrease chances of legal challenges [13]. Han *et al.* [3] considered content balancing in ensuring that their two half-length FCI assessments covered the same set of concepts. To balance content in CAT, the percentage of items to be administered from each subgroup is defined in advance (for example, to be the same as what is found in the FCI itself). Doing so ensures that items from each subgroup are administered.

Content balancing is an important aspect of instrument validity; however, implementing an algorithm which balances the content may come at a cost. In particular, when the algorithm is not used, CAT generally begins with the more informative items. When an algorithm to control content balancing is implemented, these informative items are less likely to be selected, and this may compromise the accuracy and precision of the instrument. If all items of the FCI (30 items) are administered in FCI-CAT, the accuracy and precision with/without content balancing will be identical. However, if the test length is too short, content balancing may significantly decrease the accuracy and precision.

Our goal is to analyze the effect of content balancing on the accuracy and precision of the standardized mean difference. Specifically, our research questions are: (i)How much does content balancing decrease the accuracy and precision in the standardized mean difference at a given predetermined length for the FCI-CAT? (ii)What is the minimal test length of the content-balanced FCI-CAT at which content balancing does not significantly affect the accuracy and precision?

To analyze the effect of content balancing for the FCI-CAT, we conducted a Monte Carlo simulation. Monte Carlo simulations generate responses with pseudorandom numbers, and they are commonly used in CAT development [14]. All of our analyses were conducted using R [15]. In addition to the basic package of R, the simulations of the FCI-CAT were conducted using the package CATR [16].

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 [17]. Although there are various IRT models to choose from, we use the three-parameter logistic (3PL) model to facilitate comparison with our previous study [5]. In the model, the probability of a correct response of the ith respondent on item j is given by

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

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 [5]. In order to estimate the item parameters of the FCI, the full-length paper-and-pencil (in-class) FCI was administered to 2882 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, aberrant responses were removed 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. In order to validate the model, we confirmed that the assumptions of unidimensionality, overall local independence, and goodness of fit are satisfied for the 3PL model.

B. Computerized adaptive testing

1. *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 [6]: (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 [6]. 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 our case, as is commonly done [6], we set the prior population mean proficiency value to be zero to have the scale be centered on respondents. Consequently, the first item administered to all respondents is question 13 of the FCI.

(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. The commonly chosen options for this selection are the expected *a posteriori* (EAP) method to estimate the proficiency and the MFI criterion to choose the next item. The EAP is a commonly used Bayesian method to estimate the proficiency, and, compared to several alternatives, it has been found to be less biased [18]. At this stage, content balancing can be controlled using the appropriate test algorithms as we describe below.

(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.

(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.

2. *Content balancing*

To balance the content of the items administered in CAT, 1) the items administered must be classified into subgroups and 2) the relative proportion of items from each subgroup to be administered must be specified [6]. In order to meet these requirements, we utilized the taxonomy table of the "Newtonian concepts in the revised Force Concept Inventory" [19]. Although content grouping can be determined in various ways, we chose to use this taxonomy table because we want to include all FCI items in the item pool (we did not use, for example, factor analysis [20] to determine the subgroups, because such an analysis leaves some FCI items unassigned to the subgroups). In the table, the correct responses of the inventory items are classified into six subgroups of Newtonian concepts: Kinematics, First Law, Second Law, Third Law, Superposition Principle, and Kinds of Force. Although many FCI items correspond to just one subgroup, questions 12 and 14 are assigned to two subgroups (Kinematics and Kinds of Force). Moreover, some items are also assigned to secondary subgroups. For example, question 8 is primarily assigned to the First Law subgroup, and secondarily assigned to the Second Law subgroup and the Superposition Principle subgroup. To simplify the process of separating items into respective subgroups, we consider only primary subgroups in the analysis that follows. Moreover, following the modified taxonomy table of [19] in Ref. [20], we assigned questions 12 and 14 to the Kinematics subgroup. As a result, the relative proportions of the items to be administered from each subgroup are calculated as: Kinematics (17%), First Law (27%), Second Law (10%), Third Law (13%), Superposition Principle (0%), and Kinds of Force (33%). The 0% for Superposition Principle indicates that all items assigned to this subgroup were primarily assigned to a different subgroup; consequently, we ignored this subgroup in our analysis.

There are various algorithms available to control content balancing, but the CATR package [6] allows use only of the simplest option, the constrained content balancing method [11]. The content balancing algorithm begins with the second item administered (the first item is question 13, as described above). In the case of the package CATR, the steps of the algorithm selecting the second and subsequent items are:

1. Before the administration of each item, compute the percentage of items that have already been administered from each subgroup.
2. Target the optimal subgroup of items. Generally, this is the subgroup for which the gap between the observed relative proportion of administered items and the expected relative proportion is maximal. When multiple subgroups have not yet had any of their items administered, then one of those subgroups is chosen at random.
3. Once the subgroup for the next item has been chosen in step 2, an item is selected from this subgroup to be the next item administered. As described above, the MFI criterion is used to choose the item from within the subgroup.

As described above, when the MFI criterion is used, the most informative item (question 13) is selected as the first item. This item is from the subgroup Kinds of Force. Following the algorithm above, the second item will come from a subgroup that is chosen randomly from the remaining four subgroups. After an item from each subgroup is administered, the sixth item will be chosen from Kinds of Force, since the gap between the observed relative proportion (20%) and the expected relative proportion (33%) is maximal.

C. Approach to analyzing optimal test length

As mentioned above, to analyze the optimal length of the FCI-CAT, we calculated the accuracy and precision of the standardized mean difference, Cohen's d in particular. The population parameter of Cohen's d is given by the following equation [10],

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

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 post-population (we assume that the two population standard deviations are the same, as is done in most parametric data analysis techniques [10]).

We represent the test length as l, and express the estimator for d of the test length l as $\hat{d}_l$. From within the family of estimators for d, we use the following definition for repeated measures [10],

$$\hat{d}_l = \frac{\bar{\theta}^l_{\text{post}} - \bar{\theta}^l_{\text{pre}}}{s_l}, \tag{3}$$

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 details, see Eq. (6) of Ref. [5])

In our Monte Carlo study, we generated pre- and post-responses to the FCI-CAT and, keeping class size and population parameters fixed, calculated $\hat{d}_l$ 10 000 times for each l to analyze the sampling distribution of $\hat{d}_l$. We represent the accuracy and precision in terms of the root-mean-square error (RMSE) defined by the following equation [21],

$$\text{RMSE}(\hat{d}_l) = \sqrt{E[(\hat{d}_l - d)^2]}, \tag{4}$$

where $E(x)$ is the expected value of x. With this statistic, we compared the accuracy and precision of the FCI-CAT. If the difference in RMSE values with/without content balancing is less than 1%, we regard the estimation quality of the l-length administration of the FCI-CAT to be uncompromised by content balancing.

D. Procedure of simulation study

Our Monte Carlo simulation process consisted of two steps to generate paired pre- and post-responses. 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. These were generated from designated population parameters of the bivariate normal population distributions for pre- and post-true proficiencies. 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 the empirical data previously obtained [5]. 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 (i.e., 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 post-responses and estimated proficiencies for 100 000 simulees for each l of the FCI-CAT.

Finally, 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), since we found that the RMSE depends on the class size [5]. For example, in the case of a class size of 100 students, 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.

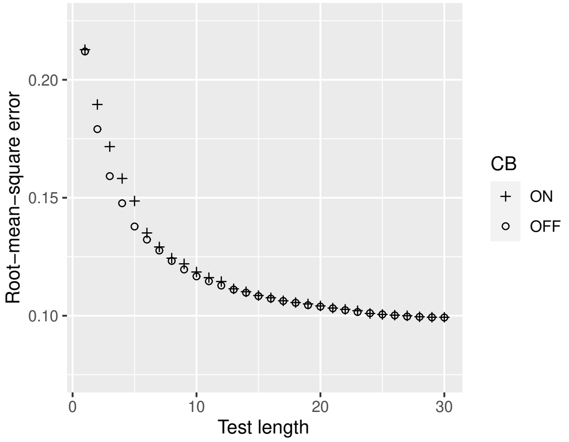

FIG. 1. The effect of content balancing on the RMSE of Cohen's d based on the FCI-CAT using the MFI criterion (class size=40). The plus symbol shows the result with content balancing (CB) and the circle symbol shows the result without content balancing.

III. RESULT

Figure 1 shows the relationship between test length and RMSE of Cohen's d based on the FCI-CAT with/without content balancing. The MFI criterion is used, and the class size is set to 40. The test length of the pre- and post-test is fixed to be the same. For the following explanation, we represent the percent difference of the RMSE between the cases with/without content balancing at each test length as $\Delta\mathrm{RMSE}(l)$. If only one item is administered ($l = 1$), the most informative item is selected regardless of whether or not the content balancing algorithm is used; thereby the RMSE differs only by statistical error ($\Delta\mathrm{RMSE} < 1\%$). When additional items are administered, the RMSE when content is balanced is larger than when content balancing is not used. The range of the $\Delta\mathrm{RMSE}$ is 6%–7% for test lengths of 2–5 items and 1%–2% for test lengths of 6–12 items. For test length larger than 13 items, the difference of the RMSE is again negligible ($\Delta\mathrm{RMSE} < 1\%$).

This result is consistent with our expectation that the differences in accuracy and precision (with/without content balancing) should diminish as the test length increases. Note, however, that there is a sudden decrease in $\Delta\mathrm{RMSE}$ at a test length of 6 items. As we described above, when the test length is 1–5 items, an item from each of the five subgroups is administered. The sixth item chosen, however, is the second most informative item in the Kinds of Force subgroup, which is the largest subgroup (the most informative item in the subgroup is administered as the first item). Since Kinds of Force is the largest subgroup, it is most probable that the item selected will be highly informative, resulting in $\Delta\mathrm{RMSE}$ dramatically decreasing.

Our result indicates that for a sufficiently large test length (say, larger than 13 items), content balancing does not compromise the accuracy and precision of the FCI-CAT, and we recommend that if one implements content balancing for the FCI-CAT, one should set the test length to be larger than 13 items. 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.

IV. CONCLUSIONS

In this paper, we considered content balancing for the FCI-CAT, 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 was defined in advance. We conducted a Monte Carlo simulation to analyze how implementing the content balancing algorithm affects the RMSE of the standardized mean difference. As a result, we found, that for a class size of 40, the increase of the RMSE due to content balancing is less than 1% for test lengths larger than 13 items. This result suggests that, if one implements content balancing for the FCI-CAT, one should set the test length to be larger than 13 items. As we discuss in Ref. [5], 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. This data consists exclusively of responses from Japanese students. Additional research is necessary to see how the results are different for other student populations.

The FCI-CAT can be validated and improved in means other than content balancing as well, for example, in terms of test security [11], removing gender unfair items [22], or by dropping one item from each locally dependent pair [23]. These concerns are important aspects to consider when thinking about instrument validity. Improving the FCI-CAT with these considerations may increase the test length required to achieve $\Delta\mathrm{RMSE} < 1\%$. Regarding validity of the test form itself, Nissen *et al.* [24] 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 data set has sufficient local independence at the whole test level [5], thereby we expect the effect of ordering is not so large. However, it is meaningful to confirm the validity of the CAT test form by conducting the FCI-CAT in real classes and comparing the result to that of the FCI-CBT.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP19K03135.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[2] B. R. Wilcox and S. J. Pollock, Investigating students' behavior and performance in online conceptual assessment, Physical Review Physics Education Research **15**, 020145 (2019).

[3] 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, Physical Review Special Topics - Physics Education Research **11**, 10112 (2015).

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

[5] J. I. Yasuda, N. Mae, M. M. Hull, and M. A. Taniguchi, Optimizing the length of computerized adaptive testing for the Force Concept Inventory, Physical Review Physics Education Research **17**, 010115 (2021).

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

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

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

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

[10] H. Cooper, L. V. Hedges, and J. C. Valentine, *Handbook of research synthesis and meta-analysis*, 2nd ed. (Russell Sage Foundation, New York, 2009).

[11] G. G. Kingsbury and A. R. Zara, Procedures for selecting items for computerized adaptive tests, Applied Measurement in Education **2**, 359 (1989).

[12] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, American Journal of Physics **78**, 1064 (2010).

[13] C. G. Kingsbury and A. R. Zara, A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests, Applied Measurement in Education **4**, 241 (1991).

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

[15] Team R Development Core, A Language and Environment for Statistical Computing (2018).

[16] D. Magis and G. Raiche, Random generation of response patterns under computerized adaptive testing with the R package catR, Journal of Statistical Software **48**, 10.18637/jss.v048.i08 (2012).

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

[18] C. DeMars, Group differences based on irt scores: Does the model matter?, Educational and Psychological Measurement **61**, 60 (2001).

[19] Table I. Newtonian Concepts in the Revised Force Concept Inventory (form 081695R), one can access the file with a password from the link "Revised Table I" of, https://www.modelinginstruction.org/effective/evaluation-instruments/ .

[20] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Physical Review Physics Education Research **14**, 010124 (2018).

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

[22] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Physical Review Physics Education Research **14**, 10103 (2018).

[23] C. S. Wallace, T. G. Chambers, and E. E. Prather, Item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set, Physical Review Physics Education Research **14**, 10149 (2018).

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