A feasibility study to develop chain computerized adaptive testing for the Force Concept Inventory

Jun-ichiro YASUDA (1), Michael M. HULL (2), Naohiro MAE (3), and Kentaro KOJIMA (4)

(1) Centre for the Studies of Higher Education, Nagoya University, Nagoya, Aichi 464-8601, Japan
(2) Department of Physics, University of Alaska Fairbanks, 1930 Yukon Dr, Fairbanks, Alaska 99775, USA
(3) Osaka University, Research Centre for Nuclear Physics, Ibaraki, Osaka 567-0047, Japan
(4) Faculty of Arts and Science, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan

Abstract. Assessment tests are commonly used to measure the pedagogical effect. To shorten the test length, we previously suggested the use of computerized adaptive testing (CAT). Based on the study, we propose increasing the frequency of CAT-based assessments during the course, while further reducing the test length per class, thus decreasing the total number of test items during the course. For that purpose, we utilize a CAT algorithm including collateral information, which we call Chain-CAT. We present the design of the algorithm, a preliminary result of analysing its efficiency by numerical simulation and discuss the feasibility of this system.

Introduction

Administering assessment tests before and after instruction is commonly used to measure the pedagogical effect. After collecting the pre- and post-test scores, we calculate a statistic such as average normalized gain and analyse the average change in students' understanding after instruction. For example, the Force Concept Inventory (FCI) [1] is an assessment test used to probe students' conceptual understanding of Newtonian mechanics. The test has 30 items with five choices, and students typically take 20 to 30 min to complete the test.

To reduce the amount of time spent on testing in the classroom, researchers have been exploring alternative approaches to test administration. One approach is to administer the assessment via online platforms which enables students to complete the test outside of class time. While this preserves in-class time, it does not reduce the amount of time that students could otherwise spend doing homework or independent study. To shorten the test time, Yasuda *et al.* [2] suggested the use of computerized adaptive testing (CAT), a practice in which a computer administers successive test items to match the current estimate of the student's proficiency.

Although CAT provides possible solutions to reduce test time, there are still specific issues associated with the pre-post paradigm. First, the current algorithm can only reduce the test length to about half without compromising accuracy and precision [2]. Second, the pre-post test results provide only snapshots of students' understanding at the beginning and end of the course, limiting the ability to observe the progression of their conceptual understanding throughout the duration of the course. Third, students may see little benefit in their own taking both pre- and post-test administered as an assessment test. This is because, in many cases, the focus of the assessment tests is to reflect on the year's instruction and improve instructional practices for future students, with no feedback to the students who take the assessment test. This situation can make students feel burdened and may decrease their engagement and motivation to take the assessment test.

To address these issues associated with the pre-post paradigm, we propose increasing the frequency of CAT-based assessments during the course, while further reducing the test length per administration, thus decreasing the total number of test items during the course. This frequent administration of short CAT-based assessments can be used as a form of formative assessment. Providing automated feedback according to each student's set of responses is expected to increase the usefulness of the survey for students and reduce their sense of burden.

The feasibility of the above ideas depends on how far the CAT-based test length per class can be reduced. For CAT where the item bank consists of the 30 FCI items, a reference value of the total test length during the course is 60 items, as this is the total test length if the full FCI is administered pre-post test. If a course involves 10 administrations of the FCI, it is then desired that the CAT-based test length per administration would be less than 6 items. To reach this goal, we utilize a CAT algorithm that takes advantage of collateral information [3]. Collateral information is the relevant empirical information on the respondents, for example, age, grade, or previous test scores. This information can be used to select the first item in CAT and to specify the prior distribution for the proficiency estimation based on the Bayesian method. In so doing, we can accelerate the convergence of the estimates during the test administration, hence improving test efficiency. Specifically, we use the proficiency estimate of each respondent in a test for selecting items and estimating respondent proficiency level in the next test. Since this CAT algorithm links the test result of each class sequentially, we call this algorithm *Chain-CAT*.

The objective of this study is to examine the feasibility of the Chain-CAT version of the FCI (ChCAT-FCI). Specifically, our research questions are: Is it possible to use the FCI Chain-CAT in a course such that the total number of items administered is less than 60 without compromising accuracy and precision? Under what conditions is this possible?

Method, Result and Conclusion

We conducted numerical simulations to analyse the efficiency of the ChCAT-FCI and to search for the optimal algorithm. The procedure is as follows. (1) Assume a progression model of true-value proficiency θ. In this study, we tried a) a stationary model in which θ is constant with respect to time, b) a linear model in which θ increases linearly, and c) a step model in which θ increases significantly only at a certain point in time. (2) Generate response data for a given true value of θ based on the item response model and calculate the estimates of θ. In this analysis, we generated 1000 response data for each time point. (3) Calculate the root-mean-square error (RMSE) from the true and the estimated values of θ and compared to a reference value (RMSE in the pre-post paper-and-pencil FCI).

To use the ChCAT-FCI as a practical test, it is necessary to add constraints to the testing algorithm. For example, it is necessary to set conditions to ensure that the items in each test are not biased toward specific concepts (e.g., Newton's Third Law), or to limit the number of times an item is repeated in different tests. We also analysed how these constraints affect the efficiency of the ChCAT-FCI in the numerical simulations.

Preliminary simulation results showed that in most cases without constraints, the total length of ChCAT-FCI in a course could be shorter than the traditional FCI testing. However, when the constraints (item balancing and/or exposure control) are imposed, preliminary results indicated that the accuracy of the ChCAT-FCI is slightly lower than that of the traditional FCI testing. The results may suggest that the CAT algorithm needs to be improved and the item bank expanded beyond the FCI items.

References

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

[2] J. Yasuda, M. M. Hull, N. Mae, Improving Test Security and Efficiency of Computerized Adaptive Testing for the Force Concept Inventory, *Phys. Rev. Phy.s Educ. Res.* **18** (2022) 010112.

[3] W. J. Van Der Linden, Empirical Initialization of the Trait Estimator in Adaptive Testing, *Appl. Psych. Meas.* **23** (1999) 21.