Preliminary Investigation of Validating Chain Computer Adaptive Testing Based on the Force Concept Inventory

Haruko UEMATSU (1) , Taku NAKAMURA (2), Michael M. HULL (3), Kentaro KOJIMA (4),
Naohiro MAE (5) and Jun-ichiro YASUDA (6)

(1) Department of Physics, Tokyo Gakugei University, Koganei, Tokyo 184-8501, Japan
(2) Department of Physics, Gifu University,Yanagito, Gifu, Gifu 501-1112, Japan
(3) Department of Physics, University of Alaska Fairbanks, 1930 Yukon Dr, Fairbanks, Alaska 99775, USA
(4) Faculty of Arts and Science, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan
(5) Osaka University, Research Center for Nuclear Physics, Ibaraki, Osaka 567-0047, Japan
(6) Centre for the Studies of Higher Education, Nagoya University, Nagoya, Aichi 464-8601, Japan

Abstract. This study explores the feasibility of Chain-CAT, a Computer Adaptive Testing (CAT) approach integrated into the pre-post assessment paradigm in educational contexts. We propose increasing CAT frequency while shortening per-test duration and reducing the total number of items. Utilizing collateral information in CAT algorithms, specifically Bayesian-based proficiency estimation, facilitates efficient testing. A preliminary investigation involving FCI-CAT implementation and interviews suggests potential for Chain-CAT to accurately measure Newtonian mechanical thinking and aid in assessing conceptual understanding progression.

Introduction

In educational settings, pre- and post-assessment tests are commonly employed to gauge the efficacy of instructional interventions. The Force Concept Inventory (FCI) [1] serves as an assessment tool to measure conceptual understanding in Newtonian mechanics. To streamline testing procedures, Yasuda et al. suggested the use of Computer Adaptive Testing (CAT) [2], wherein test items are dynamically administered based on the student's estimated proficiency level. However, integrating CAT into the pre-post paradigm presents distinct challenges. Firstly, there's a desire to further reduce test length without compromising accuracy and precision, yet current algorithms only achieve about a half-length reduction for both pre and post-tests. Secondly, the results of pre and post-tests only provide snapshots of student understanding at the beginning and end of a course, limiting the ability to observe conceptual understanding progression throughout the course duration. Thirdly, students may perceive little benefit in taking both pre and post-tests as the focus is often on retrospective review for instructional improvement rather than providing feedback to students. In such scenarios, student motivation and engagement in the assessment process may diminish.

To address these issues associated with the pre-post paradigm, we propose increasing the frequency of CAT-based assessments throughout the course while further shortening the test duration per session and reducing the total number of test items. Frequent CAT aligns well with formative assessment practices and offering automated feedback based on individual performance in each assessment could enhance the perceived utility for students, thereby reducing their burden and improving engagement.

The feasibility of the above idea depends on how far the CAT-based per-test length can be reduced. To achieve this, we use a CAT algorithm that includes collateral information [3]. This information can be used to select the first item in the CAT and specify a prior distribution for the Bayesian-based proficiency estimation. Doing so can speed up the convergence of estimates during test implementation and improve test efficiency. Specifically, the proficiency estimate for each respondent on one test is used to select items and estimate the respondent's proficiency on the next test. Since this CAT algorithm links the test results of each class sequentially, we refer to this algorithm as Chain-CAT.

The objective of this research is to validate the Chain-CAT version of the FCI. Specifically, we seek to answer whether Chain FCI-CAT can measure Newtonian mechanical thinking with accuracy and precision akin to FCI, and if it can aid in assessing students' conceptual understanding progression and formative evaluation. Herein, we report on a preliminary investigation to assess the validity of Chain FCI-CAT.

Method

To evaluate the validity of Chain FCI-CAT, we employ microgenetic methods [4] wherein students' proficiency is assessed through interviews conducted after each assessment test. This preliminary study aims to establish methods for interviews and speech analysis. It encompasses five sessions within one course, involved three participants selected considering their FCI pre-test scores, gender, and major. The research methodology included conducting physics classes integrating student interactions, covering content aligned with FCI. Following each class, participants completed a 10-question FCI-CAT test online, followed by interviews conducted in a think-aloud format. Interviews focused on the reasoning behind the choice of answers for the first five questions of each FCI-CAT test. No follow-up questions were posed during the interviews, and feedback was not provided to students through FCI-CAT or interviews. Post-research, participants were surveyed about their experience with the study.

Discussion

It was found that each subject had key questions in the FCI-CAT whose answers changed over the duration of the physics course, yet not all were addressed in this preliminary study. Interviewing on questions where responses changed could yield more participant insights. Both FCI-CAT response times and interview remarks decreased over sessions, likely due to participant adaptation. Insufficient participant remarks may hinder analysis. While follow-up questions were omitted this time, it's crucial to emphasize the think-aloud method consistently and pose follow-up queries without intervening in participant thinking as needed. We considered two evaluation methods, one that counts all correct and incorrect statements made by the respondent, and another that forms a subjective holistic impression based upon the overall interview. Although the “counting” method is more objective, it may be influenced by participant response frequency. To enhance the objectivity of the "holistic" method, we created an assessment rubric.

These preliminary findings lay the groundwork for further exploration into the viability of Chain FCI-CAT as a robust assessment tool for tracking conceptual understanding and facilitating formative evaluation in educational contexts.

References

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

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

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

[4] R. Brock and K. S. Taber, The application of the microgenetic method to studies of learning in science education: characteristics of published studies, methodological issues and recommendations for future research, *Stud. Sci. Educ*. **53** (2017) 45.